Online Grocery Recommender Using Filtering

Patchipulusu Bhuvana Chandrika
Research scholar (CSE) in International School of Technology and Sciences for Womens (ISTS women's college) in NH-16, East Gonagudem, Rajanagaram, Andhra Pradesh, 533294
.E-mail: bhuvanachandrika666@gmail.com

G.Suresh
Associate professor and hod (CSE) in International School of Technology and Sciences for Womens (ISTS women's college) in NH-16, East Gonagudem, Rajanagaram, Andhra Pradesh, 533294

V. Anil Santosh
Associate professor and hod (CSE) in International School of Technology and Sciences for Womens (ISTS women's college) in NH-16, East Gonagudem, Rajanagaram, Andhra Pradesh, 533294
Email: anilsantosh1984@gmail.com.

Abstract: With the exponential increase in information, it has become imperative to design mechanisms that allow users to access what matters to them as quickly as possible. The recommendation system (RS) with information technology development is the solution, it is an intelligent system. Various types of data can be collected on items of interest to users and presented as recommendations. RS also play a very important role in e-commerce. The purpose of recommending a product is to designate the most appropriate designation for a specific product. The major challenge when recommending products is insufficient information about the products and the categories to which they belong. In this paper, we transform the product data using two methods of document representation: bag-of-words (BOW) and the neural network-based document combination known as vector-based (Doc2Vec). We propose three-criteria recommendation systems (product, package and health) for each document representation method to foster online grocery shopping, which depends on product characteristics such as composition, packaging, nutrition table, allergen, and so forth. For our evaluation, we conducted a user and expert survey. Finally, we compared the performance of these three criteria for each document representation method, discovering that the neural network-based (Doc2Vec) performs better and completely alters the results.

***Keywords*: recommender systems; retail market; digital transformation; grocery industry; bag-ofword; Doc2Vec; nutrition tab**

I INTRODUCTION

According to [1], digital transformation facilitates new ways of value creation at all stages of the consumer decision process: pre-purchase (need recognition, information search, consideration or evaluation of alternatives), the purchase (choice, ordering, payment), and the post-purchase (consumption, use, engagement, service requests). This value creation is especially relevant in retailing to ensure competitiveness and gain a larger market share. Digital transformation came hand in hand with the penetration of mobile devices and data science in e-commerce. Although digital transformation [2] has been addressed from several approaches; multi-channel solutions, user modeling, Internet of Things, and so forth; all of them rely to some extent on the availability of information on operations, supply chains and consumer and shopper behaviors. One of the imperatives in this digital transformation is obtaining a view of customer insights. From the early steps (Amazon, 2003 [3]), the time to select the desired product has been the main issue for customers, especially if the high volume and rhythm of incorporation of products are considered. From more than two decades, Recommender Systems (RS) in e-commerce have tried to provide the most suitable products of services, to mitigate the product overload problem and to narrow down the set of choices [4–6

The recommendation can be carried out with several approaches depending on the type of data collected and the ways it is used by the RS: Content-

Based (CB) filtering, Collaborative Filtering (CF), and hybrid. Both systems CB and CF are widely used, and specially the item-based collaborative filtering where the similarity between items is calculated using users' ratings of those items. (developed by Amazon [3]). Although RSs are used by users regularly in almost all digitalized sectors, its popularization in the grocery market, that is, a retail store that primarily sells food products, has been delayed as a consequence of the low penetration of online grocery shopping, the implementation of e-commerce for grocery goods. Recently, as well as in other sectors, the grocery industry is harnessing digital to innovate through data-drive business models. Online grocery is considered a central element in the new normal. In this respect, grocery recommendation uses customer's shopping history and product information to address various added value scenarios; predicting customers' future shopping, selecting best value for money products, offering new products user may like, and so forth. Besides, the availability of data about products and shopping positively affects the retailer by easing a sustainable business; offers & featured products, stock management, customer profiling, and so forth. To meet the challenges above, in this paper, we use two document representation methods—BOW and Doc2Vec—to manage product data. We also address the three-criteria recommendation systems; Product, Package and Health for each document representation model to the specific problem of, given a source product P, applying RSs to suggest similar alternative products where similarity is defined on the basis of a product taxonomy, as well as product characteristics; composition, packaging, nutrition table, allergens, etc. The solution to this problem supports various regular use cases in the grocery market, such as out of stock products, inventory clearance, best value options, new products, etc. In order to obtain the recommender model and to validate them, we use a real grocery dataset, referred to as MDD-DS, provided by Midiadia, a Spanish company that works on grocery catalogs. MDD-DS was constructed by analyzing the product's information (product labeling) and by experts' manual annotation so that products are assigned to a specific variety in a hierarchical structure for products. Therefore, the major contributions of this research work are the following: 1. Definition of an appropriate data structure to manage the different kinds of information linked to commercial products (especially in the food industry). 2. Definition and identification of the appropriate document representation that works with MDD-DS to represent the products. 3. Design and implementation of a RS that automatically provides alternative products when the user's choice is not available. The RS do not work with user's profile, it is exclusively based on the product's characteristics and the available catalogue. 4. Design of three recommendation approaches based on the product's characteristics; composition, packaging, nutritional table, allergens, etc. 5. Proof of concept and validation to test the RS performance. We have conducted a survey for users and for experts to evaluate the RS approaches. The rest of this document is organized as follows: In Section 2, we briefly reviewed RS and document representation methods to manage product data in RS. The grocery MDD-DS is describing in Section 3. In Section 4, the recommendation methodology is introduced with three specific approaches to product similarity, based on product composition, packaging, and healthy characteristics. To implement these three approaches to product similarity, we deployed two kinds of document representation techniques: a simple BOW (Bag of Words, in Section 5) and a neural network-based word embedding, Doc2Vec in Section 6. For the two product representation models, experimental evaluation and discussion are described in Section 7. Finally, in Section 8, we conclude the current work with some future research directions.

II RELATED WORK

RS are a fundamental task for e-commerce, as the personal RS recommends providing items or products that satisfy the interests of different users according to their different interests and also recommends unknown items for the users that satisfy their interests [9]. As mentioned above, the three most commonly used methods in the RS are CB filtering, CF, and hybrid approach. CB filtering [10–12] is one of the standard techniques used by RS. CB identifies items based on an analysis of the item's content, similar to items known to be of interest to the user. For example, a CB website recommendation service can work by analyzing the user's favorite web pages to generate a profile of commonly occurring terms. Then use this profile to find other web pages that include some or all of these terms. CB technique has several issues and limitations [13–15]. For example, (i) having no mechanism to assess the quality of an item supported by CB methods. Furthermore, CB methods generally require items to include some type of content that is amenable to feature extraction algorithms. As a result, CB technique tend to be ill-suited for recommending products, movies, music titles, authors, restaurants and other types of items with little or no useful and analyzable content; (ii) CB is also have another problem that they rarely reflect current user community preferences. In a technique that recommends products to users, for example, there is no mechanism to favor items that are currently “hot sellers”. Moreover, existing systems do not provide a mechanism to recognize that the user can search for a particular type or category. CF [16,17] is another common recommendation technique. In general, the CF recommends the item to the user based on a community of user interests, without any analysis of the item content. CF idea is to build a personal profile of ratings data through each item sold and rate it through the user. Besides the CF technique’s concept to recommend the item to the user, the user’s profile is initially compared with other users’ profiles to identify one or more similar users. These similar users’ highly-rated items are recommended to the user. A significant benefit of CF is that it overcomes the previously mentioned shortcomings of CB filtering. The main issue in the above is how to measure user similarity. This problem inspires memory-based methods [18], which can be implemented as user-based [19] or item based [20,21]. User and item-based methods have similar mechanisms, but item-based methods are used more to perform better at scale and with a lower rating density. A hybrid approach is an approach that combines CB and CF (user-based and item based) filtration approaches with attempts to eliminate their flaws and provides a more efficient result. It usually performs better than either filtering method alone. Here, the hybrid approach does combine the CB and CF to solve the significant problems that are the cold start [22] and sparsity problems [23]. The cold start problem occurs when there is not enough new user data or ratings for a new item, so it is difficult to make recommendations for that new user or present new items to a user. Regarding sparsity, it occurs when the user has not rated most of the items and the ratings are sparse. In our work, we have some issues in providing a recommendation service and associated methods for generating personalized items. Science, the recommendation is based on the user’s interests without considering the user profile. This paper focuses solely on the user’s interest and how to recommend suitable items to each user. The benefit of this work is also that recommended items are identified by lists of similar items to the desired item. As mentioned above, in our paper we worked on combining CB filtering and CF (item-to-item), such as Amazon [3]. Amazon invented an algorithm that began looking at items themselves. It analyzes the recommendations through the items purchased or rated by the user and matches them with similar items, using metrics and composing a list of recommendations. That algorithm is called “item-based collaborative filtering”. This approach was also very appropriate and faster, especially for huge data sets. It was developed in 2017 [24], to aggregate data about the user to develop an RS to rely on the data and the user behavior in selecting the items. It is still based only on the analysis of Sensors 2021, 21, 3747 4 of 30 the items. However, it combines the analysis of the items with the user’s data and choices. Regarding the related works, we see that the most widely used in the previous works is collaborative filtering, as shown in the following paragraphs. In [25], the authors used a collaborative filtering method to create the proposal for various items using accessible ratings and comments on Twitter. The authors have also evaluated the reviews given by blipper (a review website) for four unique products using the CF method. When dealing with video as data to find suitable items for the user, there are also research works that apply collaborative filtering to recommend products through this kind of data. For instance, in [26], the authors introduced an approach that includes itemto-item collaborative filtering to discover exciting and meaningful videos among the largescale videos. This method runs on Qizmt, which is a.NET MapReduce framework. The RS in [27] also depends on monitoring the video content the user watches, the customer carrier database, and the vector database of products; therefore, the idea is to identify an item related to a part of the video content the user viewed that, and consequently determine the product category associated with the item, then analyze the characteristics of items similar to the item. That has been identified through the video’s visualization, and it compares the customer value vectors and the product characteristics vectors. Moreover, start showing the recommended product to the customer. Other approaches take user interactions into account to recommend the right products. For instance, in [28], the recommender system collaborative filtering uses user interactions and keeps them to benefit the recommendation. It does not stop at the items that have been selected only from the users, but the proposed system is related to the category of items. Recommendation systems usually require a large amount of user data. Safeguarding the privacy of this information is an important aspect that must be taken into account. For instance, in [29], an arbitrable remote data auditing scheme is proposed. This is based on a non third-party auditor for the network storage-as-a-service paradigm. The authors have designed a network storage service system based on blockchain, in which the user and the network storage service provider will generate the integrity metadata of the corresponding original data block respectively. All of that reach a consensus on the matter by means of the use of the blockchain technique. Other approaches solve some problems in the recommendation system, such as scalability and the cold start problem. For instance, the authors of [30] implement a user-based collaborative filtering algorithm on a distributed cloud computing platform that is Hadoop to solve the scalability problem of the collaborative filtering method. Besides, the authors of [31] propose a keyword-Aware Service Recommendation method called KASR. They also present a personalized service recommendation list and keywords used to indicate user preferences. A user-based collaborative filtering algorithm is adopted to generate the recommendations. They implemented KASR on Hadoop with real-world data sets to improve its scalability and efficiency in a big data environment. Furthermore, in [32] proposed a novel approach based on item-based CF use of BERT [33] to help understand the items and work to show the connections between the items and solve problems that are related to the traditional recommender system as cold start. This experiment was performed with an actual data set large scale with a whole cold start scenario, and this approach has overtaken the popular Bi-LSTM model. It used the item title as content along with the item token to solve the cold start problem. The approach also further identifies the interests of the user. Other approaches consider recommending products that are in line with the user's interests without being affected by the problems faced by the recommendation system mentioned above and the problem of data sparsity. For instance in [34], a product recommendation system is proposed where an autoencoder based on a collaborative filtering method is employed. The experiment result shows a very low Root Mean Squared Error (RMSE) value, considering that the users' recommendations are in line with their interests and are not affected by the data sparsity problem as the datasets are very sparse. Sensors 2021, 21, 3747 5 of 30 In e-commerce, user data and purchasing behavior play an important role [35,36]. However, in our scenario we are totally agnostic about the customer behavior. The company Midiadia does not provide complete e-commerce solutions, but provides enriched catalogues to e-commerce platforms. Consequently, Midiadia has not information about the customers interactions, habits or any kind of profiling. To the best of our knowledge, no other study provides a solution to this problem (recommending a similar product) taking exclusively into account the product information: ingredients, size, packaging, health messages, allergens, etc. All this consideration without going back to the customer data, depends only on the product description, such as name, brand, ingredients, legal name, and size; likewise, other data helps to know that the product is also healthy, such as sugars, fats, carbohydrates and excluding all the contents that can cause allergies. Our proposition fills an exciting void for many e-commerce dominants. Representation Models Regarding document representation models, we provide some representation models regarding the techniques used in this paper. We start with simple techniques such as BagOf-Words, TF-IDF. First, Bag-Of-Words (a.k.a. BOW [37,38]) is a basic, popular, and most straightforward approach among all other feature extraction methods. It is used to create document representations in Natural Language Processing (NLP) [39] and Information Retrieval (IR) [40]. The text is represented as a bag that contains many words. It forms a word presence feature set from all the words of an instance. The method does not care how often the word appears or the order of the words; the only thing that matters is whether the word is in the word list. It is generally used to extract features from text data in various ways. A bag of words is the presentation of text data. It specifies the frequency of words in the document. A feature generated by bag-of-words is a vector where n is the number of words in the input documents vocabulary. Second, TF-IDF [41] short for term frequency–inverse document frequency, is a technique that can be used as a weighting factor not only in IR solutions but also in text mining and user modeling. This method, as in the bag-of-words model, counts how many times a word appears in a document. However, words which are repeated so many times like the stopwords (the, of, ...) are penalized with this technique because of the inverse documentary frequency weighting. Here, the more documents a word appears in, the less relevant it is. Therefore, a word that is distinctive and frequent will be high-ranked if it appears in the query introduced by the user. On the other hand, word embedding is a term used for the representation of words for text analysis [42–45]. It also maps of words in vectors of real numbers using the neural network, the probabilistic model, or the dimension reduction on the word cooccurrence matrix. Word embeddings are also very useful in mitigating the curse of dimensionality, a recurring problem in artificial intelligence [46]. Without word embedding, the unique identifiers representing the words generate scattered data, isolated points in a vast sparse representation [47]. With word embedding, on the other hand, the space becomes much more limited in terms of dimensionality with a widely richer amount of semantic information [48]. With such numerical features, it is easier for a computer to perform different mathematical operations like matrix factorization, dot product, and so forth, which are mandatory to use shallow and deep learning techniques. Regarding word embedding, unfortunately, the representation of meaning with different symbols cannot orchestrate the same meaning as words. Early attempts solved this problem by clustering words based on the meaning of their endings and representing the words as high-dimensional spaced vectors. A new idea was recently proposed inspired by the neural network language model, and the model proposed is known as Word to Vector (word2vec) [49]. These embeddings are easy to work with since the vectors can be manipulated by many algorithms like dimensionality reduction, clustering, classification, similarity searching, and many more. Two models generate the representation of word2vec have been presented in order to produce such dense word embeddings: the Continuous Bag of Word (CBOW) model [50] Sensors 2021, 21, 3747 6 of 30 and the Skip-Gram model [51,52]. Each of the two models train a network to predict neighboring words. Suppose that a sequence of tokens (t1, . . . , tn)is provided. The CBOW model, first randomly initializes the vector of each word and then using a single layer neural network whose outcome is the vector of the predicted word, optimizes the original guesses. One can easily understand that the size of the Neural Network controls the size of the word vector. The Skip-gram model uses the word, in order to predict the context words. After explaining the meaning of Word2Vec, however, the goal of doc2vec is to create a digital representation of the document, regardless of its length. Unlike words, documents don't come in logical structures like words. In [53] they used Word2Vec template and added below paragraph id to build doc2vec. 3. Dataset The data set used in this paper was provided by Midiadia, a Spanish company which works to convert textual information in the product package into product category and product attributes by mixing automated natural language processing and manual annotation. The Midiadia DataSet (MDD-DS) is taxonomy where the 3 upper levels are called Category, Subcategory and Variety. Every product in MDD-DS includes; the taxonomy position, that is, values for Category, Subcategory and Variety as well as a set of product attributes. for example, name, ingredients, legal name, brand, product size, and so forth, as shown the extract of real data in Table 1. We have also used these product components before in [54,55] to provide a solution to automatically categorize the constantly changing products in the market, which is the first part of our investigation. • 'European Article Number' (EAN) is an internationally recognized standard that describes the barcode and numbering system used in world trade to identify a specific product that is specifically packaged and has a specific manufacturer in retail. • 'Category', 'Subcategory', and 'Variety' are a hierarchy and can be displayed by a company as catalog organization levels in the classification. The companies manufacture the products and each company has an identifying name and is listed as the brand. • In addition, there are some properties compatible with the EU regulation [56], for example, name, legal name and ingredients, as indicated in Table 2. • 'Servings' is a number that determined based on the amount of product and is sufficient for how many people. In addition, Midiadia supported us with two versions of MDD-DS to implement recommendation systems and cover all the company's requirements. The basic version which was called MDD-DS1, contained all the above information plus some information related to the nutrition table, such as sugar and fat, and some messages on the product packaging such as the sugar-free or the free gluten and other messages on the cover of the product. Of course, these messages are placed according to the components of each product.

The proposed methodology is shown in Figure 1. In order to obtain the model, a training set is defined in order to obtain the recommender model with the following steps: (A) MDD-DS is preprocessed; (B) for every product P the dataset is filtered by allergen preconditions; (C) the three similarity scores are obtained (PRO-COM, PK-BD, and HTHBD). Then at the bottom of the model is the automated recommendation when the user selects the product. The recommendation system recommends an alternative based on the three approaches. A survey is conducted to consider the users in the three approaches.

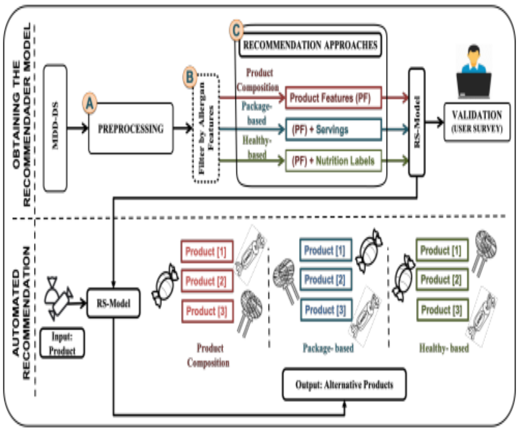

Figure 1. Description of the proposed model definition and evaluation.

III CONCLUSIONS

The recommendation idea is to implement some approaches that help a user to get the right product. The approaches are made based on the user's interest. For example, suppose the user is interested in a specific size product or a product that does not contain an allergen, and it is not available in stock. In that case, the RS recommends a similar product with these specifications without referring to the user's file; recommended depending only on the user's choice. The recommendation system can recommend an alternative health product to the user. In this paper, to build a recommendation system, we used item-based collaborative filtering (RS-CF) and BOW to represent the dataset as a vector. To build an RS-CF that caters to the largest number of users, we created three approaches, which are product composition (PRO-COM), package-based (PK-BD), and the healthy-based approach (HTH-BD). Essentially, PRO-COM works to obtain a similar product based on the product's component, whereas the PK-BD approach takes into consideration PRO-COM and adds product size to obtain a similar product. Finally, the HTH-BD approach obtains a similar product by taking PRO-COM and allergen information into account, then an equation is made, consisting of the features of the nutrition table. The user then evaluates these approaches through the survey.

REFERENCES

[1] Reinartz, W.; Wiegand, N.; Imschloss, M. The impact of digital transformation on the retailing value chain. Int. J. Res. Mark. 2019, 36. [CrossRef]

[2] Wessel, L.; Baiyere, A.; Ologeanu-Taddei, R.; Cha, J.; Blegind Jensen, T. Unpacking the Difference between Digital Transformation and IT-enabled Organizational Transformation. J. Assoc. Inf. Syst. 2020, 22, 102–129.

[3] Linden, G.; Smith, B.; York, J. Amazon. com recommendations: Item-to-item collaborative filtering. IEEE Internet Comput. 2003, 7, 76–80. [CrossRef]

[4] Thorat, P.B.; Goudar, R.; Barve, S. Survey on collaborative filtering, content-based filtering and hybrid recommendation system. Int. J. Comput. Appl. 2015, 110, 31–36.

[5] Grbovic, M.; Radosavljevic, V.; Djuric, N.; Bhamidipati, N.; Savla, J.; Bhagwan, V.; Sharp, D. E-commerce in your inbox: Product recommendations at scale. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 1809–1818.

[6] Shen, J.; Zhou, T.; Chen, L. Collaborative filtering-based recommendation system for big data. Int. J. Comput. Sci. Eng. 2020, 21, 219–225. [CrossRef]

[7] Bennett, J.; Lanning, S. The netflix prize. In Proceedings of the KDD Cup and Workshop, New York, NY, USA, 12 August 2007; Volume 2007, p. 35.

AUTHORS PROFILE

Patchipulusu Bhuvana Chandrika completed her B.Tech in 2019 in Computer Science and Engineering and has interest in Software Testing, Operating System, DBMS, Machine learning,working on deep learning techniques for Online grocery recommender system using collaborative filtering.

Mr.G.Suresh (M.Tech,Phd) he working as an associate professor in ISTS women's engineering college.

V. ANIL SANTOSH (B.Tech, M.Tech) earned his B.Tech in IT in 2005 in GIET College. He got his M.Tech in 2012 in KIET College. He published about 20 scientific papers in journals & in many IEEE & other international conferences in the area of evaluating Artificial Intelligence and Cyber Security, now working as an Associate Professor and HOD in ISTS Women's College.