Grocery Recommendation

System



Submitted to

Savita Seharawat

Submitted by

Dilpreet Singh

Ankush Kumar

Table of Contents

Abstarct…………………………………………………………………………………………………3

1. Introduction………………………………………………………………………………………...4
2. Related Work………………………………………………………………………………………5
3. Methodology…………………………………………..……………………………………….….8
4. Grocery Recommendation Datasets ……………………………………………………………...12

4.1 Data Preprocessing……………………………………………………………………..………13

* 1. Exploratory Data Analysis……………………………………………………………………..13
  2. Figure1: Most ordered products……………………………………………………….14
  3. Figure2: Number of orders taken by day of week……………………………………15
  4. Figure3: Number of orders taken by hour of the day………………………………….16
  5. Figure4: Number of products per department…………………………………………17
  6. Modelling…………………………………………………………………................................18

**Abstract**

Recommender systems are one of the most successful and widespread applications of machine learning technologies in business. You can find large-scale recommender systems in retail,video on demand, ormusic streaming services. In order to develop and maintain such systems, a company typically needs a group of experienced data scientists and engineers.

To build a recommendation system for a grocery store to find out what product a customer is going to purchase is quite complicated. In the case of this grocery store, the same customers can buy a different basket of products each time. Analyzing the cases to solve this problem in this recommendation system we cannot simply feed data in a machine learning model for the recommendation.

Now a days there is a lot of trends of online shopping. Almost all the products which are used in day-to-day life are available online especially grocery. So, there are many projects going on online grocery shopping systems. And most of them have recommendation systems in them. Recommendations are used for making the work of the customer easier and faster. This reduces their valuable time and also the efforts. For this the recommendations given to the customer should be exact and should be fast. And most importantly they should none irritate the customer. These recommendations are mostly given based on their necessity and their interest. Therefore, the customer’s necessity can be predicted from their purchase history and customer’s interest can be predicted from the people have interest same as that customer.

**Tools :** Python , Tableau

**Keywords:** KNN, SVD

**Research Questions:**

**Q1.** Analyze grocery purchase data and group like customer segments.

**Q2.** How can we personalize marketing to our customer base in unique ways?

**Q3.** How can we recommend products to our customers?

1. **Introduction**

In our project of online grocery recommendation system, we are going to develop a recommendation system which will recommend the customer products of his interest and necessity. One more additional feature in our project is that the customer is recommended a special basket based on his profile and also purchase history to some extent.

The main objective of this project is to find out which products the customers might like to purchase based on his/her previous purchase history. Recommendation systems also allows preparing more relevant personalized offers.

Recommendations are used for making the work of the customer easier and faster. This reduces their valuable time and also the efforts. For this the recommendations given to the customer by this system is exact and fast.

1. **Related Work**

In today’s society most individuals use the internet to search for information, to communicate with their social network or to buy goods and services (Ergezer, 2016). Next to this, the amount of internet users worldwide is constantly increasing and is currently around 4.5 billion, of which 2.1 billion people buy goods and services online (Clement, 2019; Internet World Stats, 2019).

Research on RSs began in the 1990s and gained more attention after the growth of e-commerce (Alyari & Navimipour, 2018). A variety of terms refer to RSs, such as interactive decision aid system (Li & Karahanna, 2015), “recommender systems, recommendation agents, shopping agents, shopping bots, and comparison-shopping agents” (Xiao & Benbasat, 2007, p. 1). The general process of RSs can be explained using Adomavicius and Tuzhilin’s three-stage process model, which is depicted in Figure 1 (Li & Karahanna, 2015). 1. In the first stage the main aim is to understand the consumer. a. Consumer information collection: the RSs implicitly or explicitly collects consumer information (Li & Karahanna, 2015; Verruck & Nique, 2017).

An example of an explicit method is using a questionnaire to elicit personality, past purchases or demographics, and an implicit method is using clickstream or social media information from the consumer or their friends. Although the explicit methods require effort from the consumer, it results in more accurate recommendations and it did not lead to dissatisfaction compared to implicit methods. However, it is best to use both methods to increase the accuracy even more (Li & Karahanna, 2015). An important aspect to take into consideration is the knowledge of the consumer. If the consumer has little knowledge about the product, service, or vendor the RS must elicit consumer’s needs and should show less recommendations (Ghasemaghaei, 2020).

However, if the consumer is more familiar with the product, service, or vendor the RS should elicit the consumer’s preferences of product attributes, such as price, brand, and reputation (Ghasemaghaei, 2020; Huang, 2016). b. Building consumer profile: a consumer profile can be built based on the selected consumer information. Online companies do not use all information since this is not optimal and realistic. According to Li & Karahanna (2015) social-network information is likely to be used more together with product attributes and consumer demographics in the future. 7 2. In the second stage the main aim is to identify and delivering recommendations. a. Matchmaking approaches: different matchmaking approaches can be used to identify the products or services that matches the consumer’s profile. The information used to build the consumer profile determines which approach will be chosen (Li & Karahanna, 2015). In section 2.1.2 ‘Matchmaking Approaches of Recommendation Systems’ different types of matchmaking approaches will be discussed in detail. b. RS presentation: the next step is the RS presentation or interface design of the recommendations.

If the design is not adequate, consumers might not understand the recommendations or ignore them. Things that need to be taken into consideration are for example, the number of recommendations shown on one page (set size), the degree to which the recommendation matches the consumer’s profile (sorting cue), recommendation instruction facilities, and whether or not to use an animated persona (avatar) (Li & Karahanna, 2015). Stage 1 and stage 2 determine the accuracy of the personalized recommendations. 3. In the third stage the impact of the RS on the consumers, companies, and market is measured. The recommendations are adjusted based on the feedback from the customer (Li & Karahanna, 2015).

Shortcomings of the Online Consumer Decision-Making Model Although this model is already very extensive, it still has a few shortcomings. First of all, the model shows a linear process, while consumers often move back and forth between the stages (Wang et al., 2016). Next to this, the model assumes that the consumer mainly uses system 2 to maximize their decision, which is rational, slow, conscious, and costs energy and effort (Chugh & Bazerman, 2009). However, most consumers make suboptimal decisions to satisfy their decision (Simon, 1972). For example, if people are hungry, emotional, or stressed, they tend to eat more unhealthy food that in the end can lead to overweight or obesity (Elsweiler, Trattner & Harvey, 2017).

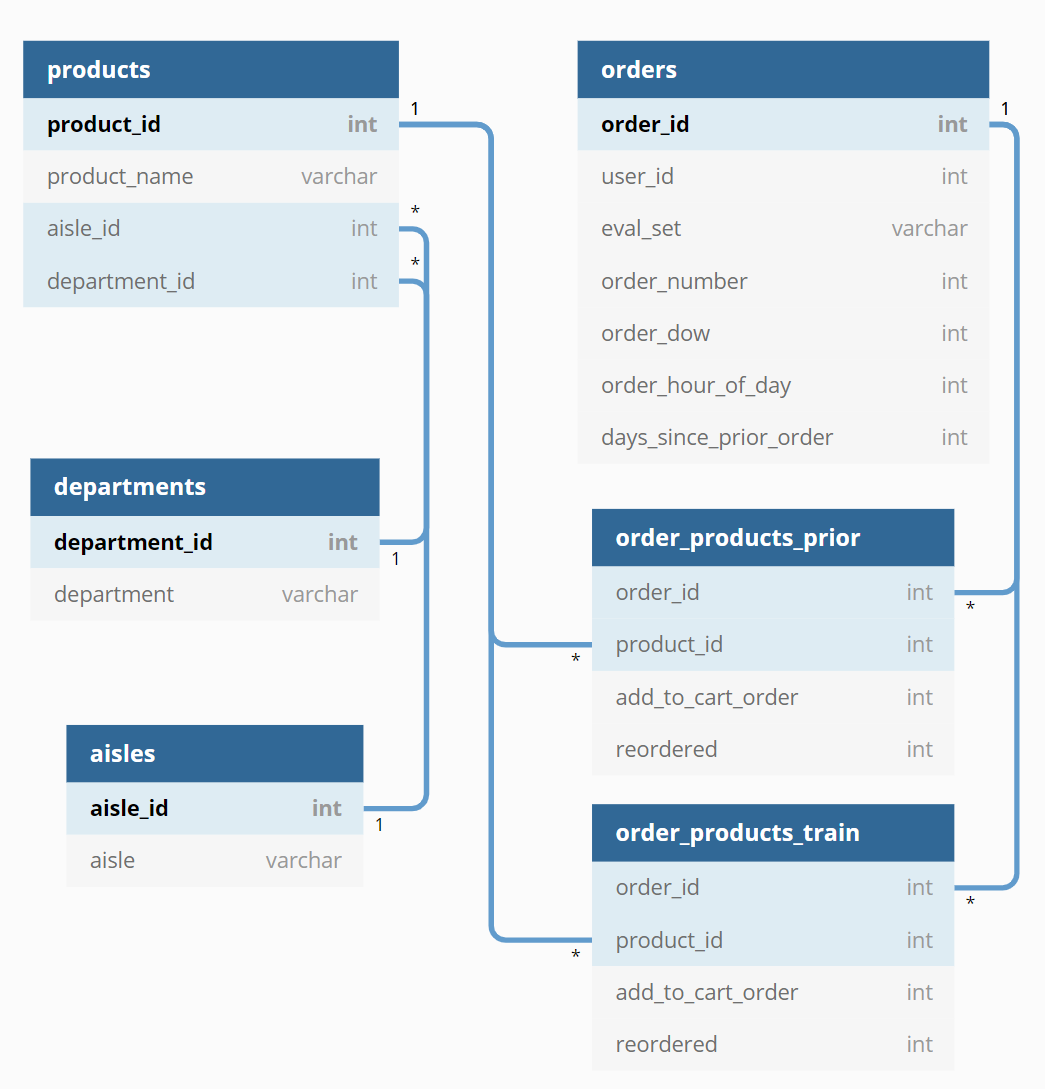
They use system 1, which is intuitive, emotional, fast, automatic, and effortless (Chugh & Bazerman, 2009). The frequent use of System 1 can be explained by the bounded rationality theory of Simon (1972). According to the theory people can only process information up to a certain limit and they are not able to calculate the optimal choice since they often lack information and/ or time (Chugh & Bazerman, 2009; Simon, 1972). This is especially true for complex decision with a lot of alternatives (Häubl & Trifts, 2000). That is why people also use heuristics, which are rules of thumb to simplify their decision-making process. For example, the Lexicographic heuristic, where the consumer ascribes a value to all attributes of a product or service and determines which one is the most important, e.g. colour or price. After that, the consumer chooses the option with the highest value on that attribute (Bettman, Johnson & Payne, 1991). Lastly, since this is a general model for the online consumer decision-making process, food related characteristics are not included

1. **Methodology**

Diagram

Description automatically generated

1. **Grocery Recommendation Datasets :**



**Aisles.csv**: 134 Unique aisle numbers and descriptions

**Departments.csv**: 21 Unique department numbers and descriptions

**Products.csv**: 49,688 Unique product ids, with description, aisle id, and department id

**Orders.csv**: 3,421,083 Unique order id, with user id, order number, order\_dow, order\_hour\_of\_day, days\_since\_prior\_order, and eval\_set indicating if the order is in train, prior, or test

21 Departments

134 Aisles

49,688 Products

206,209 Users

3,421,08 Orders

**Numerical attributes**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Table

Description automatically generated**

**Graphical user interface, application

Description automatically generated**

**Table

Description automatically generated**

**Categorical attributes**

**Department**

**A picture containing table

Description automatically generated**

**Aisle**

**Text, letter

Description automatically generated**

1. **Data Preprocessing**

Data preprocessing is the process of transforming raw data into an understandable format. Why is Data preprocessing important?

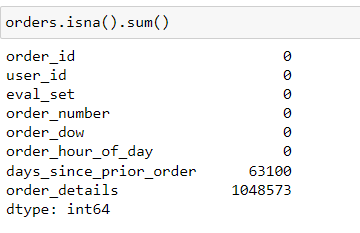
Preprocessing of data is mainly to check the data quality. The quality can be checked by the following

**Accuracy:** To check whether the data entered is correct or not.

**Completeness:** To check whether the data is available or not recorded.

**Consistency:** To check whether the same data is kept in all the places that do or do not match.

**Timeliness:** The data should be updated correctly

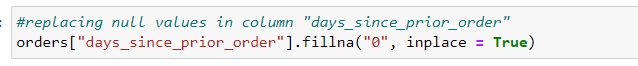
****

In order dataset we have order detail attribute with 1048573 null values and since day prior order with 63100.

Graphical user interface, text, application

Description automatically generated

In this above picture we delete the order detail attribute because it contains all null values.



In days since prior order (attribute) we replaced 0 with null values.

Text

Description automatically generated

Merging dataset:

Graphical user interface

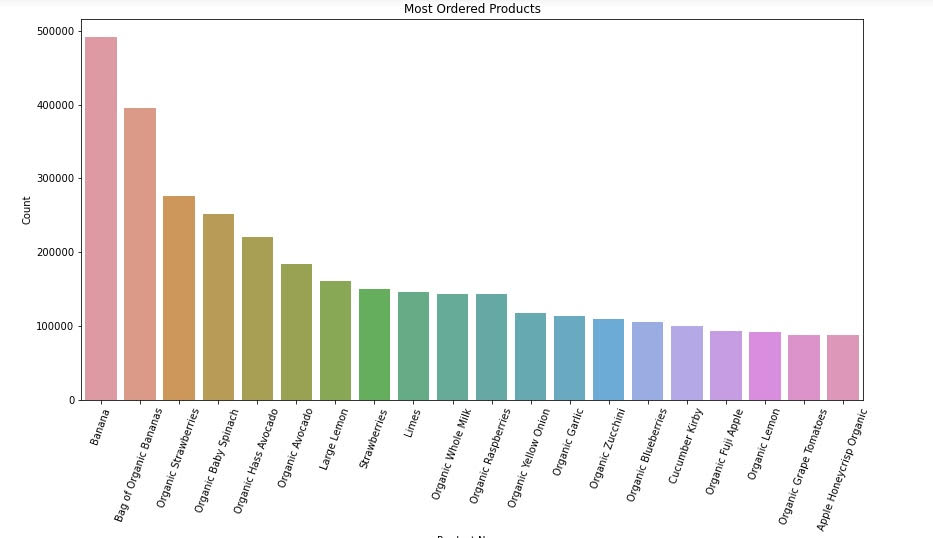
Description automatically generated with medium confidence

A picture containing table

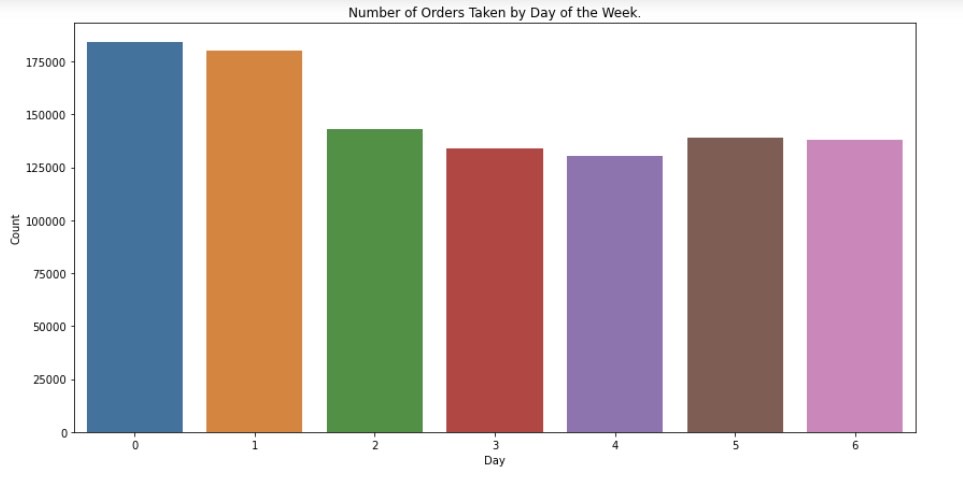
Description automatically generated

1. **Exploratory Data Analaysis:**

Useful information can be derived from just exploring the purchasing patterns in the data. We can see which aisles and departments are ordered from the most and even down to the product level. We can see the typical number of items in each order and how many days users go before their next order. As an example, the most ordered products can be seen below.

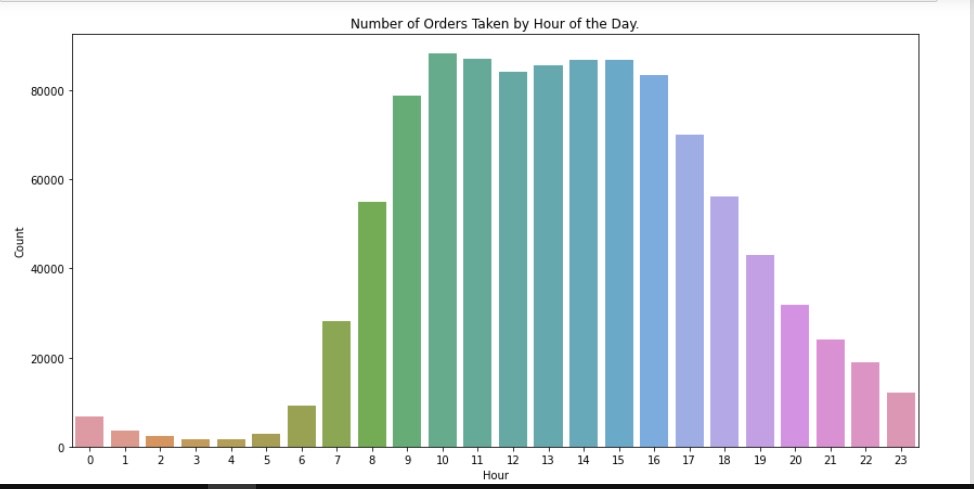
****

**Figure1: Most ordered products**

****

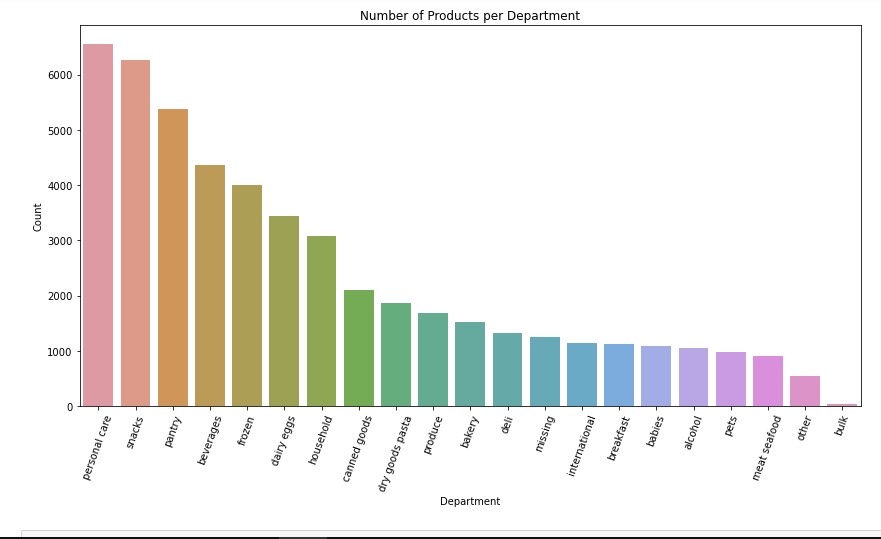
**Figure2: Number of orders taken by day of week**

**There are more than 175,000 orders on the first and second days, according to the bar graph. The third and fifth days show slight decreases in orders. Five days later, it is 125000 After the fifth day, it slightly decreases again.**

****

**Figure3: Number of orders taken by hour of the day**

**The most number of orders are taken between 10 and 16 hours of the day, as shown in the bar graph. In contrast, the least amount of orders are taken daily from 0 to 6. Furthermore, orders decrease continuously after 16 hours of the day.**

****

**Figure4: Number of products per department**

**According to the bar graph, personal care products account for the most volume per department, while bulk products account for the least amount. Additionally, snacks, pantry, beverages, frozen foods, dairy eggs and households are above 2000 based on the dataset. All other products, however, are almost exactly the same in terms of count.**

1. **Modelling:**

The size of the data, with over 32 million order id and product id combinations, was prohibitive from using a memory-based recommender, such as KNN, that would look at customer to customer similarities and item to item similarities. However, the idea of comparing similar items drove me to use Natural Language Processing to create a search engine in which one could enter any text value and get recommended products. Specifically, I tokenized and stemmed the aisle, department, and product name for each product, then using Count Vectorizer calculated a matrix of the cosine similarity for each product.

Next, I used Singular Value Decomposition, which is a matrix factorization method that finds the latent features of the customers and items while reducing the dimensionality of the data, to generate product “ratings” for each user. In the absence of explicit product ratings, I used the number of times a user purchased a particular product as a proxy for a rating, giving me a rating scale of 1–100. My initial RMSE was 3.46 which didn’t seem like a large error on a scale of 1–100 however upon further inspection I realized the items with the higher ratings had very large prediction errors.

**References:**

Zhu, D., Kuo, M. Munkhbold, E. (2016, July 10-14). Effects of E-Customer Satisfaction and E-Trust

on E-Loyalty: Mongolian Online Shopping Behavior. In T. Matsuo, A. Kanzaki, N. Komoda, A.

Hiramatsu (Eds.), Proceedings of the 5th IIAI International Congress on Advanced Applied

Informatics (IIAI-AAI) (pp. 847 – 852). Retrieved from

<https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7557730>

Müller, M., Mika, S., Harvey, M. & Elsweiler, D. (2012, May 21-24)\*. Estimating nutrition values for internet recipes. In R. Arriago & A. Matic (Eds.), Proceedings of the 6th International

Conference on Pervasive computing technologies for healthcare (pp. 191–192). Retrieved from <http://dx.doi.org/10.4108/icst.pervasivehealth.2012.248682> Nag, N., Pandey, V. & Jain, R. (2017, October 23)\*. Live personalized nutrition recommendation engine. In Proceedings of the 2nd International Workshop on MMHealth (pp. 61–68).

Retrieved from <https://doi.org/10.1145/3132635.3132643> Ng, Y.-K. & Jin, M. (2017, November 7-9)\*. Personalized recipe recommendations for toddlers based

on nutrient intake and food preferences. In Proceedings of the 9th International Conference

on Management of Digital EcoSystems (pp. 243–250) Retrieved from

<https://doi.org/10.1145/3167020.3167057> Nouh, R. M., Lee, H.-H., Lee, W.-J. & Lee, J.-D. (2019)\*. A Smart Recommender Based on Hybrid

Learning Methods for Personal Well-Being Services. Sensors, 19(2), 431:1-431:21. Retrieved

from <https://doi.org/10.3390/s19020431> Ntalaperas, D., Bothos, E., Perakis, K., Magoutas, B., Mentzas, G. (2015, October 1-3)\*. DISYS: An

Intelligent System for Personalized Nutritional Recommendations in Restaurants. In N. N.

Karanikolas, D. Akoumianakis, M. Nikolaidou, D. Vergados, M. Xenos (Eds.), Proceedings of the 19th PCI (pp. 382-387). Retrieved from

<https://dl.acm.org/doi/pdf/10.1145/2801948.2801997> Nyati, U., Rawat, S., Gupta, D., Aggrawal, N. & Arora, A. (2019). Characterize ingredient network for

recipe suggestion. International Journal of Information Technology, 1-8. Retrieved from

<https://doi.org/10.1007/s41870-019-00277-y> 48

Ohata, T., Nishihara, Y. & Yamanishi, R. (2019, June 10)\*. Foods Recommendation System for Meals- out in Nutrient Balance. In Proceedings of the 11th Workshop on Multimedia for CEA (pp. 17

23). Retrieved from <https://doi.org/10.1145/3326458.3326926> Oke, A., Kamolshotiros, P., Popoola, O. Y., Ajagbe, M. A. & Olujobi, O. J. (2016). Consumer Behavior

towards Decision Making and Loyalty to Particular Brands. International Review of

Management and Marketing, 6(S4), 43-52. Retrieved from

<https://www.researchgate.net/publication/303789700> \_Consumer\_Behavior\_towards\_Decisi

on\_Making\_and\_Loyalty\_to\_Particular\_Brands Pedersen, P. E. (2000). Behavioral Effects of Using Software Agents for Product and Merchant

Brokering: An Experimental Study of Consumer Decision-Making. International Journal of

Electronic Commerce, 5(1), 125-141. Retrieved from

<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.222> Pratibha & Kaur, P.D. (2019)\*. A Context-Aware Recommender Engine for Smart Kitchen. In Panigrahi

B., Trivedi M., Mishra K., Tiwari S., Singh P. (Eds.), Proceedings of ICSICCS AISC, vol 670 (pp.

161-170). Retrieved from <https://link.springer.com/chapter/10.1007/978-981-10-8971-8> \_16

Qiu, L. & Benbasat, I. (2010). A study of demographic embodiments of product recommendation

agents in electronic commerce. Journal Human-Computer Studies, 68(10), 669-688. Retrieved

from <https://doi.org/10.1016/j.ijhcs.2010.05.005> Rathi, S., Rander, R., Sanghvi, P. & Shah, S. (2017)\*. Eatery Directory Using Recommendation System

and Data Mining. Journal of Computer Engineering, 1, 39-42. Retrieved from

<https://pdfs.semanticscholar.org/99c6/26217fb53d543a0fdd66fc7846c1c6b8676b.pdf> Rehman, F., Khalid, O., Haq, N. u., Khan, A. u. R., Bilal, K. & Madani, S. A. (2017)\*. Diet-Right: A Smart

Food Recommendation System. KSII Transactions on Internet and Information Systems, 11(6),

2910-2925. Retrieved from <https://doi.org/10.3837/tiis.2017.06.006> Ribeiro D., Ribeiro J., Vasconcelos M.J.M., Vieira E.F. & de Barros A.C. (2018, April 28-29)\*. SousChef:

Improved Meal Recommender System for Portuguese Older Adults. In: Röcker C., O’Donoghue

J., Ziefle M., Maciaszek L., Molloy W. (Eds), Proceedings of ICT4AWE 2017, vol 869 (pp. 107

126). Retrieved from <https://doi.org/10.1007/978-3-319-93644-4_6> Saito, K., Asada, T., Yoshitomi, Y., Kato, R. & Tabuse, M. (2018)\*. A Recipe Decision Support System

Using Knowledge Information and Agent. Journal of Robotics Networking and Artificial Life,

5(3), 204-207. Retrieved from <https://www.researchgate.net/publication/331484364> Salloum, G. K. (2018)\*. Automated knowledge-based nutrition health assessment, recommendation,

progress evaluation, and meal planning (Master Thesis, Lebanese American University,

Beiroet, Libanon). Retrieved from <https://laur.lau.edu.lb:8443/xmlui/handle/10725/10468> Sano, N., Machino, N., Yada, K. & Suzuki, T. (2015)\*. Recommendation system for grocery store

49

considering data sparsity. Procedia Computer Science, 60, 1406-1413. Retrieved from

<https://doi.org/10.1016/j.procs.2015.08.216> Schäffer, H., Hors-Fraile, S., Karumur, R. P., Valez A. C., Said, A., Torkamaan, H., Ulmer T. & Trattner,

C. (2017, July 2-5). Towards Health (Aware) Recommender Systems. In P. Kostkova, F. Grasso,

C. Castillo, Y. Mejova & A. Bosman (Eds.), Proceedings of the 2017 International Conference

on Digital Health (DH ’17) (pp. 157-161). Retrieved form

<https://doi.org/10.1145/3079452.3079499> Siddaway, A. P., Wood, A. M. & Hedges, L. V. (2019). How to do a Systematic Review: a Best Practice

Guide for Conducting and Reporting Narrative Reviews, Meta-Analyses, and Meta-Syntheses.

Annual Review of Psychology, 70, 747-770. Retrieved from https://doi.org/10.1146/annurev

psych-010418-102803

Simon, H. A. (1972). Theories of Bounded Rationality. In C. B. McGuire & R. Radner (Eds.), Decision

and Organization (1st ed., pp. 161-176). Amsterdam: North-Holland

Publishing Company. Retrieved from

<http://innovbfa.viabloga.com/files/Herbert> \_Simon\_\_\_theories\_of\_bounded\_rationality\_\_\_1

972.pdf Skjold, K., Øynes, M., Bach, K., Aamodt, A. (2017, June 26-28)\*. IntelliMeal - Enhancing Creativity by

Reusing Domain Knowledge in the Adaptation Process. In Proceedings of the ICCBR, vol. 2028

(pp. 275-282). Retrieved from

<http://www.iccbr.org/iccbr17/ICCBR2017> \_workshop\_proceedings.pdf

Snyder, H. (2019). Literature review as a research methodology: An overview and guideline. Journal

of Business Research, 104, 333-339. Retrieved from

<https://doi.org/10.1016/j.jbusres.2019.07.039> Subramaniyaswamy, V., Manogaran, G., Logesh, R., Vijayakumar, V., Chilamkurti, N., Malathi, D. &

Senthilselvan, N. (2019)\*.

store. In Proceedings of the 5th international conference on IUI (pp. 260–263). Retrieved

from <https://doi.org/10.1145/325737.325866> Talekar, P. S., Raghavendra, G. S. & Vaddatti, B. (2019)\*. NutriSmart- Food Products

Recommendation System. International Journal of Innovative Science and Research

Technology, 4(5), 773-775. Retrieved from <https://ijisrt.com/wp> content/uploads/2019/06/IJISRT19MY474.pdf Teng, C.-Y., Lin, Y.-R. & Adamic, L. A. (2012, June)\*. Recipe recommendation using ingredient

50