

DYNAMIC INFERENCE OF PERSONAL PREFERENCE FOR NEXT-TO-PURCHASE ITEMS BY USING ONLINE SHOPPING DATA

Yun-Rui Li¹, Ting-Kai Hwang² and Shi-Chung Chang¹

¹Department of Electrical Engineering,
National Taiwan University, Taipei, Taiwan

²Department of Journalism, Ming Chuan University, Taipei, Taiwan

ABSTRACT

With more and more people shopping online, companies deal with customer data input not only in high volume but also dynamic. In order to attract target customers more effectively and to provide customers with more personalized services, how to automatically extract personal preference from the real-time data and make real-time recommendation has been growing in importance for businesses in the competitive modern society. Current data analysis methods for online shopping recommendation largely rely on historical transaction record. Analyses have indicated that next items a customer would like to buy not only depend on one's past historical records but on the item currently being put into the shopping cart. This paper designs an engine to combine each customer's past transaction and current shopping cart data to dynamically infer one's preference for the next items. The design, Transaction-Data Based Real-time Preference Inference Engine (TRPIE), consists of two innovative ideas. The first exploits the purchasing sequence information and turns one's purchase history into a temporal series of data, where a customer's dynamic purchasing behaviour information lies. The second is a design of a two-layer Recurrent Neural Network (RNN) for extracting personal purchasing preference pattern from the temporal series of data to infer preference of next items. A reference implementation of TRPIE design integrates existing tools such as Keras, tensorflowTM, sklearnTM, and MlxtendTM. Test results over real data from 1,374 people show that prediction accuracy has doubled that obtained by a basket analysis method, which ignores sequentiality of purchasing items.

KEYWORDS

Online shopping, transaction data, purchasing sequence, dynamic preference inference, recurrent neural network.

1. INTRODUCTION

In order to keep pace with customers, it is imperative for companies to be able to predict customer behavior such that they can cater to the needs of customers in the future [14]. However, customer needs, situations, expectations, and demands constantly change and evolve [21]. So there would be no way to understand them beyond 'today' without some way of predicting customer behaviour. Predicting future customer behavior has thus been an elusive goal. But, today, with the advancements in data analytics, predicting customer behavior has become more achievable than ever before.

In practice, it's very common to speculate customers' future behavior by predicting individual customer preferences. With customer preferences predicted, retailers may, for instance, entice extra purchases to the target customers. In a study of customer preference [3], the author made a survey on product rankings in order to figure out customer preferences.

In the era of e-commerce, real-time marketing allows brands to engage with their customers based on real-time information, such as their behavior and actions on a website. More specifically, real-time marketing needs to measure customer behavior in real time for adjusting companies' marketing strategies based on their holistic shopping behavior and to offer more relevant services that are individualized and contextual. With digital transformation, automated data collections and increase of computing power, predictive analytics software reliably helps forecast and influence purchasing behavior [18] and more and more companies are adopting it.

E-commerce retailers have a significant advantage over brick-and-mortar businesses when it comes to collecting data about customers' transaction data. Whenever a sale transaction is made, all purchases by a customer are on record through the membership window with information items such as payment type, products purchased, date of purchased, delivery information, etc [17]. Therefore, for online stores, it is rather easy to collect the data. According to [17], customers preferences may be extracted from their high-frequency purchasing behavior. Online groceries are consumable commodities, which will be purchased repetitively and frequently by customers.

In this paper, the problem scenario is on online groceries, where the products belong to daily necessities, are of low-price and frequent and repetitive purchases in specific categories. Online grocery shopping data provides personal behavior information through both historical and current shopping cart data. The requirements for a shopping cart are that the system may collect data real-time from each shopping cart about what products a customer adds into the shopping cart. Such requirements hold in online grocery stores or checkout-free grocery stores, where real-time data is collected via various sensors, computer vision recognition and internet of things technologies..

Figure 1 depicts the specific problem scenario of this paper, where the sequentiality of personal purchasing behaviour and historical and dynamic shopping cart data input are highlighted. Instead of analysing only the statistics from historical data, this paper will, this paper will address how to combine sequential correlation among each customer's implicit and habitual purchasing behind the historical data with current online shopping cart data so as to dynamically infer one's preference for the next-to-purchase items.

This paper will use Instacart Online Grocery Shopping Dataset1 in 2017 as the test data set. Instacart is an American company that operates as a same-day grocery delivery service in the United States [12]. Customers select groceries through a web application from various retailers and delivered by a personal shipper. Instacart makes Instacart Online Grocery Shopping Dataset accessible by the general public, where there are 206,209 customers' transaction history including 3,421,083 orders in total, and it categorizes 49,685 products into 21 departments and 134 aisles.

This paper will design an engine to combine each customer's past transaction and current shopping cart data to dynamically infer one's preference for next-to-purchase items. The design, Transaction-Data Based Real-time Preference Inference Engine (TRPIE), consists of two innovative ideas. The first exploits the purchasing sequence information and turns one's purchase history into a temporal series of data, where a customer's dynamic purchasing behaviour information lies. The idea is motivated by the fact that a customer's purchasing behavior is a dynamic process [16], and the purchasing has correlation in sequence by personal purchasing habit or preference. Current basket analysis [20] does not consider such dynamic aspect. The second is a design of a two-layer Recurrent Neural Network (RNN) for extracting personal

purchasing preference pattern from the temporal series of data to infer preference of next items. A reference implementation of TRPIE design integrates existing tools such as Keras, tensorflowTM, sklearnTM, and MlxtendTM. Test results over real data from 1,374 people show that prediction accuracy has doubled that obtained by a basket analysis method, which ignores sequentiality of purchasing items. more specifically, from an average accuracy of 18.17% to 36.46% by TRPIE.

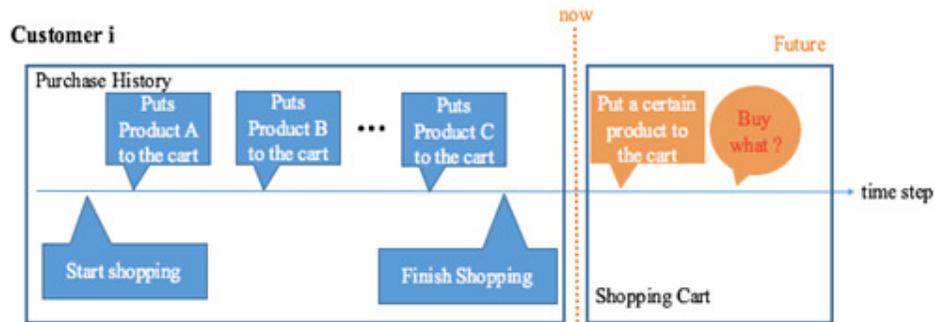


Figure1: Illustration of Problem Scenario

2. DATA ANALYTICS OF CLUSTERING FOR CUSTOMER SEGMENTATION

According to [6], even though we may have data from millions of customers, these customers may only belong to a few segments: customers are similar within each segment but different across segments. For instance, different market segments may have different product preferences and behavioral patterns. Therefore, we may often want to analyze each segment separately, as segments may have different habitual characteristics.

Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately. Segmentation allows marketers to better tailor their marketing efforts to various audience subsets. Those efforts can relate to both communications and product development. Specifically, segmentation helps a company: (1) focusing on the most profitable customers [2], (2) better predict customer's purchasing behavior for upsell and cross-sell3 other products and services [19].

In general, in order to develop refined segment-specific insights, it is often necessary to split our data into segments and perform any subsequent analysis within each segment [6]. Nowadays, clustering techniques are frequently used to do so and do well.

2.1. Data Analytics of Association Rule Learning for Discovering the Connections Between Products

According to [20], association rule learning helps a company: (1) to market precisely (e.g. target customers who buy flour with offers on eggs, to encourage them to spend more on their shopping basket), (2) to drive recommendation engines. Shortly, it not only improves the customer shopping experience but also a company's marketing campaign.

Association rule learning is a rule-based machine learning method for discovering the associations and connections between products in large-scale transaction data [1]. General speaking, market basket analysis is perhaps the most famous example of association rule, and is employed today in many application areas including web usage mining, intrusion detection, and continuous production, etc [1]. In a market basket analysis, marketer always want to see if there

are combinations of products that frequently co-occur in transactions [20]. For example, maybe people who buy flour and sugar, also tend to buy eggs because a high proportion of them are planning on baking a cake. However, unlike sequence mining, association rule learning typically does not consider the order of items within transaction data [1].

2.2. Data Analytics of Sequence Mining for Finding Statistically Relevant Patterns Between Data

Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. Furthermore, with modern technologies, it enables us to record sequences of online user activity at an unprecedented scale. Although these methodologies mentioned above have been successfully applied in many tasks, it does not take into account the temporal aspect that characterizes sequential data [6]. Specifically, those machine learning models are not suited for sequential data, since they consider each input sample independent from previous ones.

In contrast, given an input sequence, Recurrent Neural Networks (RNNs) are able to process each element at a certain time step, storing the necessary information of each element [15]. RNNs solve various problem with inherently sequential information by connecting the hidden layer with itself [6]. They are able to maintain an internal state which allows them to exhibit dynamic temporal behavior and use the hidden state as an internal memory. Shortly, RNNs process sequences in a natural way.

3. PERSONAL PURCHASING PREFERENCE PATTERN EXTRACTION

In this section, we shall proceed to the Personal Purchasing Preference Pattern (PPPP) extraction method. The goal of this section is to describe how the recurrent neural network (RNNs) is applied to PPPP extraction based on the purchase history data.

3.1. Preparation for Personal Purchasing Preference Pattern Extraction

In this subsection, we will show preliminary work for purchasing preference pattern extraction. Firstly, we show the input of purchasing preference pattern extraction. Secondly, we introduce preparation of training data with target via CPHT algorithm.

3.1.1 Input Data

We extract the following information from the purchase history data, which will be the inputs of purchasing preference pattern extraction:

- i1) User ID: It identifies customers, because we want to capture the personal purchasing preference pattern.
- i2) Order ID: It consists of Product ID, which tells us which products are bought at this order.
- i3) Product ID: To know what products a customer had purchase.
- i4) Catalog of Products: It's for the one-hot encoding³ of the specific Product ID.
- i5) Add to Cart Order: Order in which each product was added to cart. It's for the RNN to run through the sequence of items consumed by a user, item by item.

Sequential Product Sequence:

Each person's purchase history data can be represented as $\{o^t\}_{t=1}^{n_i-1}$, where o^t is the Order ID at t^{th} . And, the each order o^t within purchase history denotes as product sequences $\{x_j\}_{j=1}^T$, where x is product ID and T is Add to Cart Order.

Here, we shall give a definition of Personal Purchasing Preference Pattern (PPPP).

Definition of Personal Purchasing Preference Pattern:

According to [9], purchasing patterns reflect how customers purchase goods or service. Like basket analysis, purchasing preference pattern is like a rule that tells us when someone buy product, which product he/she will buy next. In general, personal purchasing pattern like the rule represents a person's purchasing habit or purchasing experience. Also, we focus on customer-oriented strategy and critical marketing, so use individual data to extract pattern. Then, we can predict future behavior via understanding the personal purchasing pattern. Therefore, in the research, we define personal purchasing preference pattern as two-layers RNN model and weights, $fi(\cdot)$.

Definition of Personal Purchasing Preference Pattern Extraction:

How to extract the two-layers RNN model and weights characterizing purchasing preference pattern of each customer based on the above inputs i1) to i5)?

Given a customer i 's purchase history data, $\{o^t\}_{t=1}^{n_i-1}$, design function CPHT(\cdot): {all orders within purchase history data} \rightarrow {input sequences} and {corresponding targets}, called training data with targets, so that RNN(\cdot): {input sequences} and {corresponding targets} \rightarrow $fi(\cdot)$ can fit all training data with targets to the model, $fi(\cdot)$. More detailed illustrations such as CPHT(\cdot) and RNN(\cdot) will be addressed in the following.

3.1.2 Data Pre-processing

Now, as we already know, we need the training data with target for follow-up RNNs. So, we need an algorithm to split each order within purchase history data into input sequences and corresponding targets as training data with target, such that the RNNs model can learn to predict products. Because functionality of this algorithm is converting each order within purchase history data into training data with target, we call this algorithm CPHT.

Table 1 below is an example showing the input and output of CPHT(\cdot). CPHT(\cdot) can read multiple orders at the same time and outputs the input sequences and corresponding targets of each order. The interpretation of the input sequences and corresponding targets is one by one explained in Figure 2. Each red circle in Figure 2 is called one example of training data.

Table 1. Input and output from the CPHT(.)

Input:
Each order within purchase history data, o^i $\cdot o^i : (x_1, \dots, x_T)$
Output:
<i>{input sequence} and {corresponding targets},</i> that's usually called training data with targets - {input sequence} : $\{ (x_1), (x_1, x_2) \dots, (x_1, x_2, \dots, x_{T-1}) \}$ - {corresponding targets} : $\{ x_2, \dots, x_T \}$

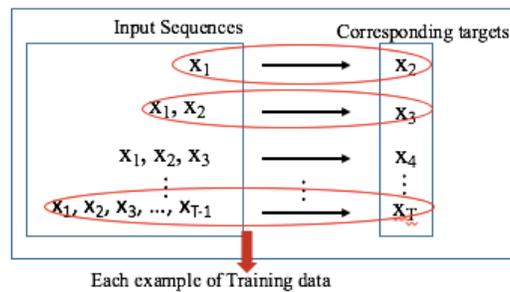


Figure 2. Exemplary of input sequences and corresponding targets from CPHT(.)

Basically, the cores of CPHT are two rules. The first rule is that if there is only one product in this order, the order will not be treated as an example of training data because we cannot know what he will buy after he has bought the product. The second rule is that the last product in each order will not be considered as an example of training. data for the same reason mentioned in rule1.

3.2. Preference Extraction by using RNNs to Exploit Sequential Information Availability

3.2.1 Recurrent Neural Networks

Recurrent neural network is one type of Neural Network architecture: It process sequences by iterating through the sequence elements and maintaining a "state" containing information relative to what they have seen so far [4]. Each person has the individual purchasing behavior process and his historical transaction record can be modeled as sequential data, which has the particular trait that can incorporate a temporal aspect [10]. Then, the process gradually becomes habits or experience [19], representing habitual purchasing behavior. As a result, RNN is naturally applied to extraction of sequential purchasing preference pattern from purchase history data for real-time marketing.

3.2.2 Long short-term Memory (LSTM)

The LSTM model was developed by Hochreiter and Schmidhuber. Primarily, it is for overcoming the problem of vanishing gradients in RNN [11]. The gradient of some of the weights in RNN would become too small or too large if the network is unfolded for too many time steps. This is called the vanishing gradients problem [7]. However, this LSTM model resembles a standard recurrent neural network with a hidden layer, but each ordinary node in the hidden layer is replaced by a memory cell [11] to solve the problem. The number of repeating modules in LSTM is determined by the length of time steps. In short, LSTM is known to learn problems with long-term temporal dependencies, so an LSTM may succeed in the sequence learning.

3.2.3 Supervised Learning and RNNs with LSTM

Supervised learning is by far the dominant form of deep learning today, with a wide range of industry applications, such as sequence generation, syntax tree prediction and object detection, etc. Generally, almost all applications of deep learning that are getting the spotlight these days belong in this scenario.

With supervised learning, humans would give input data and corresponding targets to the computer. After the computation outputs the rules, then these rules could apply new data to produce original targets [4]. Supervised learning mainly could be classed into regression task and classification task depends on the different scenarios. In our research, our task belongs to classification task, and we choose RNNs as our classifier. Because, as mentioned in the early subsection, it is suitable for handling sequential data. For handling the long-term dependencies in the sequences well, our research chooses LSTM model for our RNN to extract purchasing preference pattern from purchase history data.

3.2.4 Purchasing Preference Pattern Extraction

Table 2 below shows input and output of RNN(.) with LSTM; RNN(.) with LSTM can be fed into all examples of training data relevant to our classification task. The following covers what our RNN with LSTM looks like and how to fit all these examples obtained from CPHT(.) to a model via RNN(.) with LSTM.

Table 2. Input and output from RNN(.) with LSTM

Input:
All training data with targets.
Output:
two-layers RNN model and weights, $f_i(.)$, representing customer i 's purchasing preference pattern.

We need to understand how the RNN with LSTM processes these examples more clearly, so a step-by-step demonstration is presented through an exemplary training data with targets. RNN with LSTM processes these examples by iterating through the sequence elements and maintaining a "state" containing information relative to what they have seen so far [4].

To conclude, the RNN(.) would come up with model architecture that you assigned and trained weights for automating our classification task after seeing all examples from personal purchase history data. After training, we obtain RNN model with two-layers LSTM and weights for predicting personal preference for next-to-purchase product, $f_i(.)$. It represents customer i 's purchasing preference pattern in our research.

3.3. Learning Process of RNN

Given the input sequences and corresponding targets, a new method is required to infer preference. The learning process to assess a personal RNN classifier (as shown in Figure 3) from scratch.

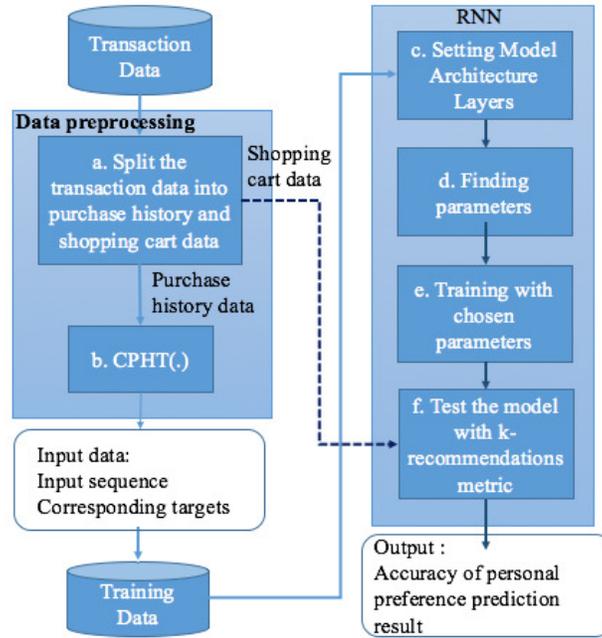


Figure 3. Learning Process of RNN

4. TRANSACTION-DATA BASED REAL-TIME PREFERENCE INFERENCE ENGINE (TRPIE) IMPLEMENTATION AND EXPERIMENT RESULT

This section presents the reference implementation of TRPIE methodology and the performance evaluation results of TRPIE over transaction data that Instacart provides.

4.1. Definition of TRPIE System

To implement TRPIE, there will be a definition of personal recommendation for next-to-purchased product.

Definition of Personal Recommendation for Next-to-Purchase Product:

Given the product that was put into shopping cart by the customer i , $p_{shopping}$, design function $TRPIE(\cdot): p_{shopping} \rightarrow \{m \text{ products that customer } i \text{ is likely to buy the next}\}$.

4.2. Architecture of Reference Implementation of TRPIE

Based on different methodologies and algorithms presented in previous sections, this section illustrates how these different parts are realized in an integrated system.

The entire system is implemented on Unix-like operating system and programmed in Python 2.7. In this implementation, scikit-learn¹ is adopted to implement clustering techniques. MLxtend² is adopted to implement market basket analysis [13]. Keras³ and Tensorflow⁴ is adopted to implement RNN-Based PPPP. All of these existing software are executed in Unix-like operating system. Within the Unix-like system, a Python environment is required to install. Software including scikit-learn¹, MLxtend², Keras³, and Tensorflow⁴ are glued together by using python scripts. In Python[®], CPHT, GPIE, PPPP extraction, and other Python scripts are built for

processing and analyzing the data are developed. To modularize, a file folder is built. The below 4 shows the mutual relationships between different systems and stacks.

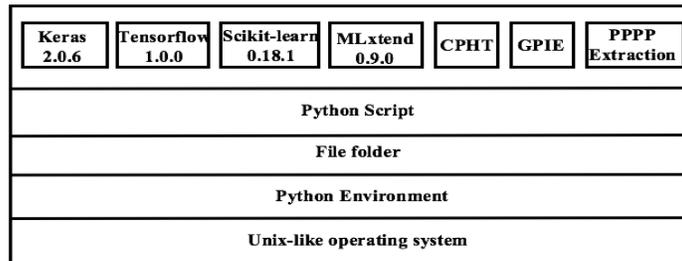


Figure 4. Software stack of TRPIE

4.3. Experimental Validation of TRPIE

The experiment is made based on Instacart Online Grocery Dataset [8]. To obtain statistically meaningful results [4] and more data to enhance power of machine learning-based methods [4], the customers who have purchased less than 99 times during data record period are filtered. At last, leaving 1,374 customers selected and to be analyzed.

Input:

We take data of 100 orders from each customer of these 1,374 customers. We consider the top 97 orders as purchase history data for training and consider the rest 3 orders will be considered as shopping cart data for testing.

Hypothesis:

Because customer's purchasing behavior is a dynamic process, we hypothesized that the RNN-based Personal Purchasing Preference Pattern considering order of purchasing items is able to better reflect dynamic purchasing behavior than something that does not consider. Therefore, we adopt preference for next-to-purchased purchase product inferred from PPPP, given the products that customers put into shopping cart, as prediction on next purchase product.

Baseline - Apriori algorithm [13]:

To the best of our knowledge, there is no comparable personal preference extraction algorithms that exploit purchasing sequence information in the literature. We shall consider the Apriori algorithm as the baseline for performance comparison with PPPP. The reason is that existing Apriori algorithm does not consider the sequential process in customer's transaction data, instead PPPP using RNN considered the sequential process.

Setting for Comparison:

We use the same purchase history data as what PPPP did. That is, the top 97 orders as purchase history data from those 1,374 people, and we uses the top 10 frequent products obtained from individual purchase history data as preference inference through Apriori algorithm (minimum support = 0.1 by default). It's often used in practical applications.

Evaluation Metric:

In this experiment, we use a short-term prediction metric to test the performance of PPPP, which refers to [5]. A short-term prediction aims to predict which product the customer i will purchase the next (i.e. right after the last one). The accuracy of short-term prediction for the performance measurement of PPPP is mathematically defined as follows.

Definition of Accuracy of Short-term Prediction:

$$\text{Individual prediction accuracy by PPPP} \equiv \frac{|P_i \cap \{x_i\}|}{|P_i|} \quad (1)$$

Result:

The numerical results are shown in Table 3. PPPP achieves 18.29% higher Individual Prediction Accuracy than Baseline on average and standard deviation is maintained at a similar scale.

Table 3. Statistics of PPPP/ Comparable Baseline Accuracy

Inference Accuracy of 1374 customers	PPPP	Baseline
Mean	36.46%	18.17%
Standard Deviation	20.66%	17.98%
Min	0.00%	0.00%
25%	22.04%	5.00%
50%	31.86%	13.63%
75%	45.45%	25.71%
Max	100.00%	100.00%

5. CONCLUSIONS

This research combines the data of the customer's past transaction record and the items currently being put into the shopping cart to dynamically infer one's preference for the next items. A transaction-data based real-time preference inference engine (TRPIE) was implemented. Also, a two-layer Recurrent Neural Network (RNN) was adopted for extracting personal purchasing preference pattern and inferring preference of the next-item purchase. In addition, the inference ability of TRPIE was demonstrated by experiment over transaction data that Instacart provided.

TRPIE was implemented by integrating different designs of methods and prestigious open-source toolkits such as Keras, tensorflowTM, sklearnTM and MlxtendTM. In the experiment, the results are as expected. PPPP, considering information of shopping cart data and dynamic customer behaviors, achieves 18.29% higher than benchmark on average.

To conclude, this study focused on the research of preference inference. Since the dynamic feature of a customer's purchasing behavior, the item of purchasing has correlation in sequence by personal purchasing habit or preference. Therefore, firms or manufacturers can fine-tune their marketing strategies in real time based on the identified customer preferences.

ACKNOWLEDGEMENTS

This work was supported in part by the Ministry of Science and Technology, Taiwan, ROC, under Grants MOST-106-2221-E-002-129 and MOST-107-2221-E-002-184.

REFERENCES

- [1] Agrawal, Rakesh. Imieliński, Tomasz & Swami, Arun (1993) “Mining association rules between sets of items in large databases”, Proceedings of the 1993 ACM SIGMOD international conference on Management of data.
- [2] Bowes, Pitney. “Customer Segmentation and Profitability”, [Online] Available at <http://www.pbinsight.com/files/resource-library/resource-files/ppb-cust-segment-and-profit.pdf>
- [3] Cao, Yu “Marketing research report of Burberry cosmetics”, [Online] Available at http://www.academia.edu/6610615/Marketing_Research_Report_of_Burberry_Cosmetics/
- [4] Chollet, Francois (2017) Deep Learning with Python (1th Edition), Manning Publications.
- [5] Devooght, Robin & Bersini, Hugus (2017) “Collaborative filtering with Recurrent Neural Network,” arXiv:1608.07400v2 [cs.IR].
- [6] Evgeniou, Theos. “Introduction to data analytics for business”, [Online] Available at <http://github.com/InseadDataAnalytics/INSEADAnalytics/tree/master/CourseSessions/Sessions45/>
- [7] Gamboa, John Cristian Borges (2017) “Deep Learning for Time-Series Analysis”, arXiv:1701.01887v1 [cs.LG].
- [8] Instacart (2017) “Three Million orders, open sourced”, [Online] Available at <https://www.instacart.com/datasets/grocery-shopping-2017/>
- [9] Kahn, Barbara (2012) “Buying Pattern”, [Online] Available at <http://kwhs.wharton.upenn.edu/term/buying-patterns/>
- [10] Ko, Young-Jun. Maystre, Lucas & Grossglauer, Matthias (2016) “Collaborative Recurrent Neural Networks for Dynamic Recommendation System”, JMLR: Workshop and Conference Proceedings 63:366–381.
- [11] Lipton, Zachary C. Berkowitz, John & Elkan, Charles (2015) “A critical review of Recurrent Neural Network for sequence learning”, arXiv:1506.00019v4 [cs.LG].
- [12] Manjoo, Farhad (2015) "Instacart's Bet on Online Grocery Shopping", The New York Times. [Online] Available at <https://www.nytimes.com/2015/04/30/technology/personaltech/instacarts-bet-on-online-grocery-shopping.html>
- [13] Mlxtend (machine learning extensions), [Online] Available at <http://rasbt.github.io/mlxtend/>
- [14] Newman, Emily (2016) “Importance of predicting customer behaviour”, [Online] Available at <http://corp.yonyx.com/customer-service/17790/>
- [15] Santolaya, Daniel Sánchez (2017) “Using recurrent neural network to predict customer behavior from interaction data”, Master Thesis, University of Amsterdam.
- [16] Solomon, Michael R. (2016) Consumer Behavior: Buying, Having, and Being (12th Edition), Pearson Education Limited.
- [17] Teng, Hui-Ping (2017) “Infer Individual Customer Preference for a New Product Based on Supermarket Transaction History,” Master Thesis, IIE, NTU.
- [18] Waxer, Cindy (2010) “What will your customers buy next”, MIT Technology Review, [Online] Available at <https://www.technologyreview.com/s/421928/what-will-your-customers-buy-next/>

- [19] Wong, Danny (2017) “How Ecommerce Companies Can Cross-Sell and Upsell to Increase Customer Lifetime Value”, [Online] Available at <https://conversio.com/deep-dive/cross-sell-upsell-increase-customer-lifetime-value/>

- [20] Yali (2017) “Market basket analysis identifying products and content”, [Online] Available at <https://discourse.snowplowanalytics.com/t/market-basket-analysis-identifying-products-and-content-that-go-well-together/1132/>

- [21] Yeong, BinCho. Yoon, HoCho & Soung HieKim (2005) “Mining changes in customer buying behavior for collaborative recommendations”, International Journal of Expert Systems with Applications, Volume 28 Issue 2, pp. 359-369.