The Review of Socionetwork Strategies (2021) 15:429–453
https://doi.org/10.1007/s12626-021-00092-7

ARTICLE

A Latent Topic Analysis and Visualization Framework for Category-Level Target Promotion in the Supermarket

Yi Sun[1] · **Teruaki Hayashi**[1] · **Yukio Ohsawa**[1]

Received: 19 May 2021 / Accepted: 10 September 2021 / Published online: 27 September 2021

Abstract

Deciding when and which products to recommend to whom is always an essential issue for retailers. In this study, we propose a mixed framework with two components to capture customer buying behavior and its changes over time and visualize these results to better help retailers choose and target products strategically for marketing. In this framework, a topic model is first used to extract customer's purchase behavior instead of association rules or K-means as mainly used in market field. To automatically choose the optimal number of topics, we implement an approach proposed by Koltcov et al. on point-of-sale (POS) data in the supermarket. Meanwhile, to grasp the change of topics over time, we divided monthly POS data in half and applied the topic model with Renyi entropy separately. The results suggest that splitting data might be a better way to understand customer behavior. Second, we consider how to develop an effective way to visualize the results of the topic model, which is essential, because in a supermarket context, simply knowing which product categories are included under which topics is not enough to support how a supermarket promotes their products. To address this, we design a three-layer visualization approach to better interpret the topic model results and to help retailers design target promotion strategies. The design of visualization was overlooked by studies related to the use of topic models on supermarket data. Finally, to demonstrate the usefulness of our proposed framework, we conduct a simple scenario-based analysis between our framework and other models, such as Latent Dirichlet Allocation (LDA) and the Dynamic Topic Model (DTM). The results show that for most periods, our proposed framework outperforms LDA and DTM.

Keywords Topic model · Renyi Entropy · POS data · Topic changes over time · Visualization

✉ Yi Sun
sun-yi650@g.ecc.u-tokyo.ac.jp

[1] Department of Systems Innovation, School of Engineering, The University of Tokyo, Tokyo, Japan

1 Introduction

How to optimize targeted promotion is always a challenge in the field of marketing. To address this challenge, some pioneering research has focused on how to build user profiles [1, 2], while others have focused on the combinations between products [3–5]. Compared to other industries, for the supermarket, because of its highly competitive and low-margin nature [6], how to utilize their data is a critical issue. In this area, prior studies have suggested several different options to help supermarkets with targeted promotion, such as predicting change points in POS data to forecast future trends [7]; rearranging the layout of the supermarket to increase impulse purchases [8–10] and categorizing customer's moving paths to decipher their decision processes [11], for example. A loyalty program plays an essential role in managing the relationship between customers and supermarkets among all these options. However, since most stores launch such programs to ensure customer retention, this program will instead become a redundant resource [12]. Therefore, it becomes essential to differentiate loyalty programs from other stores through further customized, targeted promotion.

In this research, to reach such differentiation, we focused more on the fundamentals of the loyalty program, such as deciding what products should be promoted and when, based on the analysis of customers' purchase behavior, than on how to improve the design of the program. Many researchers have also devoted themselves to this field, thereby contributing to the development of efficient algorithms which can be used on POS data. Many used market basket analyses, such as association rules [1, 4] and K-means clustering [10], for example. Using these methods, patterns of customers' purchase behavior can be extracted. One of the increasing trends in this field is to adopt topic models, a soft clustering method that can analyze and explore latent patterns in high-dimension data [13]. Thus, instead of association rules or K-means, mainly adopted by prior studies, we implement the topic model in this study.

In the study of topic models, selecting the optimal number of topics when applying the topic model is a fundamental issue [22]. Especially when using supermarket data, the number of topics representing customer's purchase tendency will directly impact understanding customer behavior and the decisions about which products should be promoted. While most of the topic models require the user to select this number manually, Koltcov's research about the application of Renyi entropy on topic models provides an automated, time-saving solution [14, 15]. Therefore, this study also implements this method to automatically search for the optimal number of topics in the POS data.

The main objective of this study is to propose a mixed framework with two parts to capture customer buying behavior and its changes over time and to interpret these results for retailers. Each of these is new in this type of study. In the first component, this study uses a combination of the topic model and Renyi Entropy with renormalization analysis to extract customers' latent patterns in the POS data. This study is the first application of this approach to POS data. Second, compared to most prior studies on the application of topic models on supermarket

POS data, they only provide text-based results of the topic model, such as using a table to list the top n word in each topic and the top m topic in documents, we address a three-layer visualization approach to explain these results better and to support retailer's decisions on target promotion. Instead of using the complete purchase history of each customer as a document, we focus more on treating each transaction data as a document to accommodate better circumstances where the customer's ID information is not available. In this way, we get a whole picture of customer's purchase tendencies. In addition, we also estimate our results and compare them with other models to explore the potential advantages of our framework.

We organize the remainder of this paper as follows: In the next section, we review the relevant literature and explain in detail the differences between this study and prior studies. Then, we describe a stepwise procedure of our frameworks and the methodology in Section 3. Next, Section 4 includes the empirical experiment we did on five-month data and presents the results in Section 5. Finally, we discuss the managerial implications and limitations of this research in Section 6.

2 Literature review

2.1 Topic Models in Marketing

These days, as the size of the data changes dramatically, a growing number of researchers in marketing are exploring the application of machine learning in this field. One application uses the topic model on marketing data. The topic model is an unsupervised machine learning method that can provide a flexible framework to soft cluster large datasets with high dimensions [13, 16]. This model was first introduced by Blei et al. to cluster large quantities of discrete text-based data to finding latent patterns [16]. They proposed a LDA model that assumes that a topic should be a distribution over words in documents, and each document is a distribution over topics [17]. Due to the topic model's high extensibility, other researchers have developed various extension models based on LDA to face different scenarios. One extension modifies the definition of documents, topics, and words to apply to different data types, such as purchase history [18, 19] and online discussion [20], for example. In this research, we likewise consider purchase data as documents. However, unlike some prior studies, we use each transaction in POS data as a document to serve our purpose. Another aspect of prior studies focuses on enhancing the utility of the model itself. For example, Blei et al. developed the DTM to capture changes in time [21].

Of course, the topic model method also has its limitations and needs to be optimized for different scenarios. One the optimizations chooses the number of topics, which is done manually in most prior studies [26], but not in our study. The choice of the number of topics will significantly affect the effectiveness of the model [22]. Too many topics, which often represent purchase tendency in the retail industry, will hinder retailers in making efficient decisions, and retailers will miss the relevant information, especially when using real-world data. Recently, by comparing

the topic model with physical systems, due to the similarities between them, physical techniques, such as entropy calculations, can be applied to optimize the topic model. Koltcov's research first applied Renyi entropy to calculate the optimal number of topics automatically. The basic idea is that when the entropy of the number of topics (like temperature in the physical system) is at the lowest point, the system will contain the most practical information. This study adopts Koltcov's research to apply to POS data, which will be the first time this technique has been tested for its usefulness outside of text data.

2.2 Topic Models in the Supermarket

In the practical application of the topic model, some researchers have directed their attention to the purchase history of customers in stores, both online and offline. Sun et al. use customer online purchase data to analyze and categorize customers through topic models that can be used to predict customers' propensity to group purchasing behavior [18]. In the offline store study, Iwata et al. expanded the topic model to include price information in supermarkets as variables under which customer characteristics can be further demonstrated [23]. Nevertheless, applications of the topic model in supermarkets are still relatively few because the data itself are not text-based, which means there is more opportunity for exploration by researchers. In this research, not only do we consider how the topic model can be applied to real-world scenarios, but we also discuss the seasonal characteristics of supermarkets by proposing a visualization approach to help retailers better understand their decisions.

Furthermore, in the business world, it is not enough to simply show raw results, for example, the top n words in each topic and the top m topic in documents. Usually, if the topic model is applied to POS data in supermarkets, the inter-connectedness of topics, the proportion of product categories in each topic, and their combination are all substantial. Blei et al.'s study also points out this problem. They consider "making this structure (topic modeling algorithms) useful, but doing so requires careful attention to information visualization and corresponding user interfaces" [17]. Therefore, this study argues that it is not enough to propose an algorithm for the practical application of topic models on data from supermarkets; it is equally important to consider how visualization can better assist retailers in their decision-making.

3 Methodological frameworks

3.1 Overview

Prior studies provide different applications of the topic model in marketing research, such as customer profiling, and purchase prediction, for example. In addition, most of these research studies deal with the development of topic model extensions, such as sLDA [22], DTM [21], User Aware Sentiment Topic Model [24], and Visual Sentiment Topic Model [25]. Although prior studies provide mixed support for the usefulness of the topic model, some missing links still need to be further discussed in

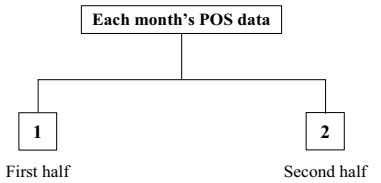
the application of the topic model on marketing data: (1) how to choose the optimal number of topics in real-world data is rarely discussed, (2) how the change of time affects the combination of product categories in the topic model, and (3) how to help marketers better understand and apply these topic model results.

Based on these missing links, the purpose of this research is to propose a mixed framework using the topic model with Renyi entropy to extract customers' purchase behavior and to design a visualization approach to interpret the results of the topic model for retailers. By observing the changes in the topic model results over time, which category should be recommended and when among the tens of thousands of categories can be determined. This purpose is accomplished by combining different data mining methods into one framework and using it on POS data. In this study, we consider each customers' transactions as a document. The product category in those transactions is a word. Thus, the topic extracted from POS is a combination of product categories, which indicate customers' purchase tendencies. In addition, we focus on presenting results to retailers, making the results easier for them to understand and deploy on-site, which is one of the significant differences from previous studies. We hope this research can lay the foundation for future studies, such as designing a new loyalty program, increasing customers' in-door experience, redesigning the supermarket's layout, and rearranging point-of-purchase displays in the supermarket, for example.

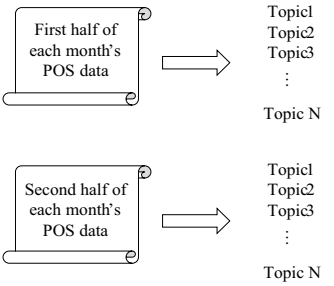
Based on prior studies, we present the framework of this research. Figure 1 illustrates the steps of our proposed framework, which uses four steps to extract latent patterns of customer purchase behavior and to discover changes in the combination of product categories in the topic model. In Step 1, to track customers' purchase behavior over time, we split each month's data in half. Next, in Step 2, we implement one of the most classic algorithms in topic models, the LDA [16] to identify latent patterns for each dataset. The number of topics in this step should be set manually with a relatively more significant number. Then, in Step 3, we use Renyi Entropy with renormalization analysis [15] to search the optimal number of topics and we use this number as input for an optimization procedure. Then, we re-run the model and extract the most valuable topics. By observing these results, lists of categories in different periods can be recommended. In the last step, we present these results by designing a three-layer visualization approach that incorporates the relations between topics, the specific items included in the product categories within the topics, and the changes in the combination of the product categories over time.

3.2 Step 1: Splitting Data

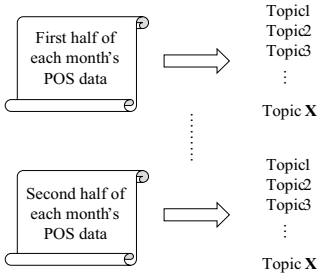
Before we implement the topic model, in this first step, we divide five-month POS data into ten datasets that allow us to capture seasonal changes of the combination of product categories in topics. In the marketing field, it is important to track changes in customers' purchase behavior over time [19]. Therefore, researchers must consider how to reflect such changes. Nevertheless, in the field of topic models, many advanced studies have discussed how to extend topic models to capture changes of words in each topic, such as DTM [21]. However, in this study, we are more inclined



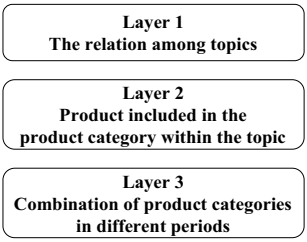
Step1. Splitting each month's data in half



Step2. For each data set, the topic model is implemented separately, and the number of topics in this step should be set manually with a relatively larger number.



Step3. Using Renyi entropy to optimize the results of the topic model and find the most appropriate number of topics , X. This number will vary with the period.



Step4. Designing a three-layer visualization framework to improve the presentation of topic model's results

Fig. 1 Steps of the proposed framework for extracting customers' purchase behavior

to divide each month's data in half and generate results separately instead of adopting such models for two reasons. First, unlike texts or articles, seasonal product categories in the supermarket will only be sold during a specific period, such as watermelon in summer or Christmas goods in winter. Thus, not only will the combination of product categories within each topic change over time, but the topic's meaning itself may continue changing over time, and such changes are critical in the study of supermarket purchases. A particular topic may disappear entirely within a certain period and reappear after a few weeks or months. However, due to the elimination of information about temporal continuity, this study also realizes that this certainly brings some drawbacks. In this study, the consistency we show is not a continuum of time. but a coherence that maintains the most effective information by finding the most appropriate number of topics. The reason for splitting each month's data in half is to reflect the impact of specific seasonal product categories in supermarkets that tend to be bought intensively in a half month, such as Christmas, Halloween, or Japanese New Year-related products.

Furthermore, with these changes, the number of topics in different periods may also be different and this information provides important managerial implications

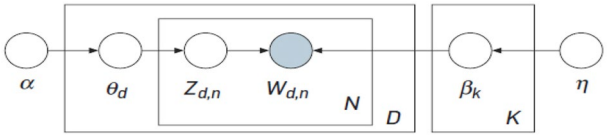


Fig. 2 Graphical model for LDA model (Blei et al. 2010)

because each topic represents a latent pattern of the customers' purchase behavior. The change in the total number of topics may reflect the interests of all customers towards the store and allow supermarkets to implement different types of target promotion strategies. In conclusion, to capture these changes, splitting data and generating results separately might be a more appropriate way to understand purchasing behavior than using models like DTM [21].

3.3 Step 2: Topic Model

This study uses the LDA model to analyze customer's purchase behavior and to set the number of topics at 100 for further optimization. Topic models have been used for analyzing text data for a very long time. As we mentioned above, recently, there is an ongoing trend for scholars to adopt such techniques in marketing [26]. For example, Iwata's research uses the topic model to analyze purchase data and cluster-related items by considering their price [23]. Before introducing this research, we introduce the concept of the topic model briefly.

Essentially, after being processed by the topic model, a document collection can generate multiple topics, in which each topic represents a distribution of words with high probability. Each document will be assigned by topics with different probabilities [27]. For example, based on prior studies, the probability of word w in document d can be expressed as follows:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \Phi_{wt}\theta_{td} \tag{1}$$

where t illustrates the topic; $p(w|t)$ is the distribution of words by topics, and $p(t|d)$ is the distribution of a document collection by topics [27].

As noted previously, over time, many extensions of the topic model have been developed. However, we use one of the most classic topic models, LDA, to analyze customers' purchase behavior in this research. Traditionally, to perform the LDA model, words in the document need to be first transferred to the bag-of-words, and the order of words in each document is ignored. Then, a set of all unique words, called a 'dictionary,' can be created using the bag-of-words. After that, by calculating distributions of words, topics can be sampled, and each distribution represents a topic. Thus, every word has a different probability of being drawn into topics. The graphical model for LDA is shown in Figure 2 [17].

In this research, every two weeks of the POS data is treated as a document collection. Each document is represented by one transaction made by a customer, including several product categories. Thus, notations in the LDA model have a

different meaning. Table 1 is a comparison table that reveals differences between a text document collection and POS data in this research.

Each of the product categories in every transaction is treated as a 'word' in one 'document,' and those topics are assigned to transactions that can be characterized as latent customer purchase behaviors in two weeks. Let d be each transaction made by customers in two weeks POS data D. And $letw_{d,n}$ represent a vector of a purchased product category in each transaction. α is a parameter ruling the per-document topic distribution that contains the distribution of words by topics. This distribution has two settings, symmetric and asymmetric. A symmetric α means each topic has the same distribution throughout the document, which is adopted by most of the prior studies and in this research.

To learn customers' purchase behavior through POS data, the LDA model is shown as follows. The transaction data contain the transaction ID, date, and product category. Each transaction d has its topic distribution θ_d that is generated by LDA, which can be described as a purchase preference. For each product category $w_{d,n}$ in transaction d, a topic $z_{d,n}$ is assigned by LDA to the transactions' topic distribution θ_d. Then, the product category $w_{d,n}$ is drawn from word distribution β_k. After this step, we can get matrix Φ that contains a distribution of words by topics, where the number of topics T is its columns' number, and the number of unique words is its number of rows. The size of this matrix is $W \times T$.

3.4 Step 3: Renyi Entropy with Renormalization Analysis

This step uses Renyi Entropy with renormalization analysis to optimize the number of topics for each period. We intend to discover the number that can demonstrate the most helpful information. Although topic models can capture latent patterns in text documents, social media, sales transactions, and more, how to optimize the result of topic models is always an important task. Among prior studies, there is a novel approach to searching the optimal number of topics by calculating and discovering the minimum of Renyi entropy of the topic model. This approach was first proposed by Koltcov's research, where the topic model's result is considered a non-equilibrium complex system, thus applying approaches from thermodynamics for quality estimation is possible [15]. The principle of this approach is to consider the topic model as a complex system that includes many "particles." This system will be in a non-equilibrium state in its initial state. The system will find equilibrium by running the topic model by exchanging energy as "temperature T" (which in the topic model is the number of topics) changes. Thus, the complex system's entropy differences can be measured to discover when an information maximum is reached. Koltcov's research considers entropy as negative information; thus, the maximum entropy corresponds to the minimum of information [14, 15]. Therefore, the value of T corresponding to the smallest entropy value can be considered the "true number of topics," representing the maximum valid information generated by the topic model.

Based on Koltcovs' research [14], the Renyi entropy can be expressed as follows:

Table 1 Comparison table for a text document collection and POS data

Notation	Description for text document collection	Description for POS data
K	Numbers of all topics	Numbers of all topics
D	Collections of text document	POS data for two weeks
W	The number of unique words in the document collection	The number of unique product categories in POS data
N	The number of words in a document collection	The number of product categories in POS data
$W_{d,n}$	n th word in document d	n th product category in transaction d
$Z_{d,n}$	Expected topics for the n th word in document d	Expected topics for the n th product category in transaction d
θ_d	The topic distribution for the d th document	The topic distribution for the d th transaction
α	Per-document topic distribution	Per-transaction topic distribution
β_k	Words distribution for a topic k	Product categories' distribution for a topic k
η	Per-topic word distribution	Per-topic product category distribution

$$S_q^R = \frac{ln\left(Z_q\right)}{q-1} = \frac{qln\left(q\tilde{P}\right) + q^{-1}ln(\tilde{\rho})}{q-1} \tag{2}$$

where $q = 1/T$ is called the deformation parameter and T is the number of topics. Z_q is the partition function of a topic solution which is shown as below:

$$Z_q = e^{-qE+S} = \rho\left(\tilde{P}\right)^q \tag{3}$$

Z_q can also be considered the statistical sum of the information system, which in this case, is the topic model. In the calculation of this partition function, E is the free energy of the topic model that can be described as follows:

$$E = -ln\left(\tilde{P}\right) = -\ln\left(\frac{1}{T}\sum_{w,t}(\varphi_{wt} \cdot 1_{\{\varphi_{wt}>1/W\}}\right)\# \tag{4}$$

where $\varphi_{wt} > 1/W$ represents words with high probabilities and $1_{\{\varphi_{wt}>1/W\}} = 1$ if $\varphi_{wt} > 1/W$ and zero otherwise. Thus, P calculates the probabilities of each word in the data by splitting the vocabulary of a given dataset. The change in the number of topics corresponds to the change of the energy in the topic model. Next, ρ represents the density-of-states function, which is formulated as:

$$\rho = N/(WT)\# \tag{5}$$

where N is the number of words with high probabilities, *and* W is the number of unique words.

In addition to the application of Renyi entropy, this study also employs a renormalization technique to accelerate the computational speed for a fast approximation of Renyi entropy and to search the optimal number of topics [14]. The procedure of implementing renormalization for the LDA model can be explained as follows:

1. A pair of topics $t1$ and $t2$ with the smallest local Renyi entropy is selected according to Eq. (2), where only the probability of words in that topic is considered.
2. Based on the matrix Φ that is calculated at step 2, we merge topic pairs into a new single topic and calculate topic's distribution. Columns in that matrix will be replaced with the distribution of words by this new topic. Therefore, the size of this matrix will be $W \times (T-1)$.
3. After new topic solutions (matrix Φ) are generated, the global Renyi entropy for this solution, which represents distributions of all topics, is calculated according to Eq. (2).

This procedure will be repeated until only two topics remain. Then, to study the behavior of the obtained global Renyi entropy and to find its global minimum, a curve of entropy as a function of the number of topics was plotted [14].

3.5 Step 4: Visualize The Result of The Topic Model

In this step, we design a three-layer visualization approach to better present the topic model results. As we discussed previously, the topic model can reveal latent patterns from data. However, if researchers only use text-based information to convey the result of the topic model, it might be challenging to make it meaningful. Simply knowing which product categories are included under which topic is not enough to support marketing by the store's retailers. Thus, how to develop an effective way to visualize and present the result of the topic model is essential in this research for three reasons.

First, based on the result of the topic model, understanding the relationship among topics is also essential, especially if we want to apply the topic model in the supermarket's context. Since the "word" in each topic for POS data represents a product category, naturally, we assume topics for POS data should have a stronger relationship with each other and be more meaningful than topics for text-based data. The reason is that words may have different meanings for different sequences in documents which is ignored because of the topic model's process (transition of bag-of-words). However, the supermarket's product categories have meaning that will not be changed by the disappearance of sequences in POS data. To visualize these relations, we adopt the method developed by Sievert and Shirley called LDAvis [28].

Moreover, this research aims to propose a framework server as a decision support tool that can support retailers' target promotions in the supermarket. To achieve this, this framework must be meaningful to those working in the field. In the supermarket, whether it is the decision of which product should be put into the loyalty program or whether the supermarket's layout should be redesigned, most promotions target specific products instead of categories. Thus, even though this research focuses on extracting helpful information at the product category level, the visualization of the result must consider the product level. Only a few prior studies about applications of topic model in marketing research have made such an effort.

Furthermore, to reveal how the combination of product categories changes over time by visualization is also vital for the analysis in the supermarket. Prior studies proposed DTM to capture topic changes over time, but as we explained earlier, this result may not be enough for the supermarket. Moreover, as the monthly data are split in half, the combination of product categories will also change over time and observing these changes may provide crucial managerial implications to the supermarket.

4 Empirical Application

4.1 Data Description and Processing

In this section, we first present in detail real-world data from one supermarket chain store in Japan that apply to the framework proposed in this paper to extract topic information. This store is in one of the sub-centers of Tokyo, surrounded by both commercial and living areas.

Table 2 Department names and the total number of types of product categories and SKUs

Names of departments	The total number of types of product categories	The total number of types of SKUs
Fruit	107	606
Vegetables	88	801
Japanese-style daily foods	69	966
Daily goods	64	1868
Meat	53	736
General foods	51	3256
Western-style daily food	49	2264
Pastry	46	3067
Fresh fish	37	1032
Side dish	36	948
Alcohol	17	1344
Concessionary•Side dish	3	28
Special•General merchandise	1	1
Special•General foods	1	1
Concessionary•Meat	1	19
Concessionary•Fruit	1	14
Clothing	1	1
Cigarettes	1	116
Bakery	1	2

The POS data used in this study from this supermarket contain more than 580,000 transactions made within five months from September 2020 to January 2021. These transactions consist of 17,070 stock keeping units (SKUs), which represent a distinct type of item in the supermarket and are the elements of the product categories. For each transaction, it contains the transaction ID, transaction date, transaction time, department, product category, product name, quantity, and revenue. There are 627 different product categories from 19 departments identified by this supermarket chain, which is shown in Table 2. These POS data are the latest data available from the supermarkets.

To apply the topic model, the POS data were converted to a form containing only three columns: transaction ID, transaction day, and product category, which is the document in the topic model. The average number of product categories is 7.9 for each transaction in five months. Figure 3 shows the conversion from POS to documents. Although information about quantity and amount is removed from the data, if a category is purchased more than once, that category will repeatedly appear in the transaction. Therefore, based on the first step of the framework we proposed, splitting the monthly POS data into the first and second halves is necessary to better understand customers' purchasing behavior and changes. Accordingly, by splitting five-month data, we obtain ten different datasets, and each product category is a "word" in documents. Therefore, each transaction represents a "document" in the datasets.

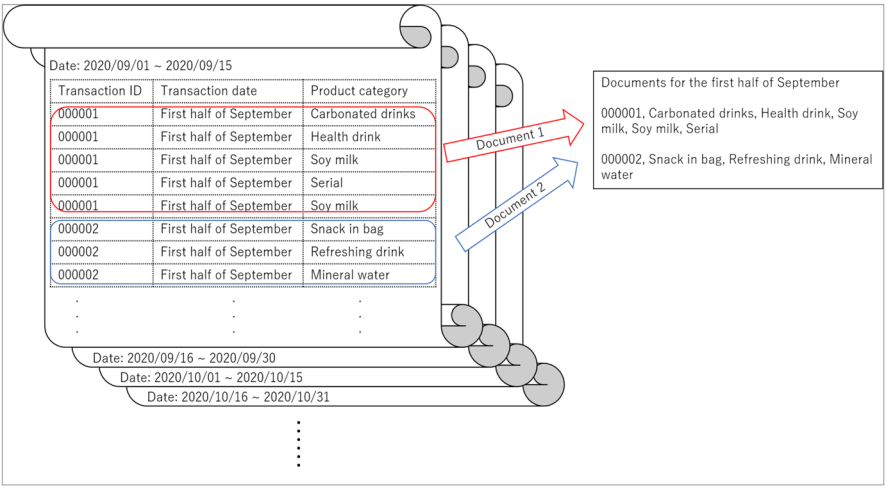


Fig. 3 The conversion from POS to documents

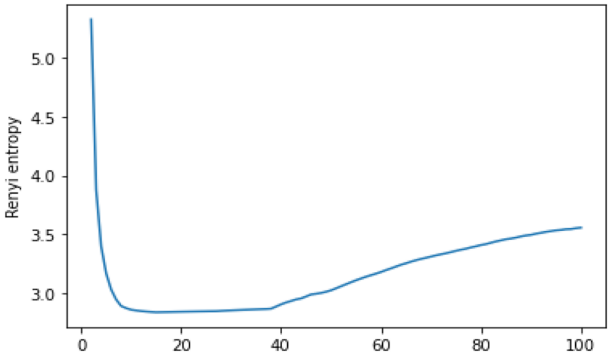


Fig. 4 Renyi entropy's curve for the POS data from the first half of September 2020. The horizontal axis is the number of topics

4.2 Step 2 and Step 3 Conducting The Topic Model with Renyi Entropy

After processing the data, we next extract a relatively more significant number of topics by running the LDA model on POS data from September 2020 to January 2021. In this study, this number is set at 100. Then, these topics from the model are used to discover the optimal number of topics by adopting the method proposed by Koltcov and Ignatenko [14]. Based on this result, a curve of Renyi entropy as the function of the number of topics was plotted. For example, Fig. 4 shows the Renyi entropy curve obtained using POS data from the first half of September 2020. The minimum Renyi entropy is located at $T = 15$, indicating that this would be the optimal number of topics that capture customer behaviors the best for this data. After

Table 3 Total number of product categories, number of topics, and the average number of product categories

Period	Number of product categories in topics	Number of topics	The average number of product categories
2020/09/01	114	15	7.6
2020/09/16	138	19	7.3
2020/10/01	102	13	7.8
2020/10/16	107	13	8.2
2020/11/01	217	30	7.2
2020/11/16	241	33	7.3
2020/12/01	240	33	7.3
2020/12/16	242	33	7.3
2021/01/01	129	17	7.6
2021/01/16	257	36	7.1

that, we extract the 15 topics in this period. Then, based on the average number of product categories, we slightly increased this number to account for future changes and set each topic to include the ten highest probability product categories.

5 Results and Discussion

In this section, we present the details of the results of this study and show the design for the visualization. Then, the estimation procedure of this research will be discussed. This procedure illustrates the potential gains of using the proposed framework.

5.1 Results of Topic Model

As we showed in section 4.2, we used the same method for other datasets to obtain the best number of topics for each period. To better show this result, each 1st of the month represents the first half of the month, and the 16th of the month indicates the second half of each month. Next, we use Table 3 to present the total number of product categories and the number of topics generated by our proposed framework. For most periods, the number of topics is well summarized under 35, representing a concentration of adequate information. Based on the results of Table 3, we discover that a higher number of topics can be considered as fragmentation of customer behavior during the period, while the opposite represents a concentration.

Meanwhile, as Table 4 shows, some product categories appear in more than one topic in the same period, such as OM2 meat, milk, cut salad, sweetened bun, etc. Moreover, we also discover that this kind of product category appears in different periods. Based on this observation, we can assume that the proposed framework captures some specific product categories with higher purchase frequency. At the same time, this result is also consistent with the general impression in the supermarket context.

Table 4 The five product categories with the highest number of topics in different periods

SKUs	2020/09/01	2020/09/16	2020/10/01	2020/10/16	2020/11/01	2020/11/16	2020/12/01	2020/12/16	2021/01/01	2021/01/16
OM2 meat	7	10	6	5	13	15	14	15	6	14
Milk	5	7	5	6	10	4	14	13	6	11
Cut salad	2	5	4	1	9	1	8	9	1	12
Pastry	4	1	5	6	3	1	12	1	8	8
Bag snack	1	5	3	2	2	1	2	10	3	9

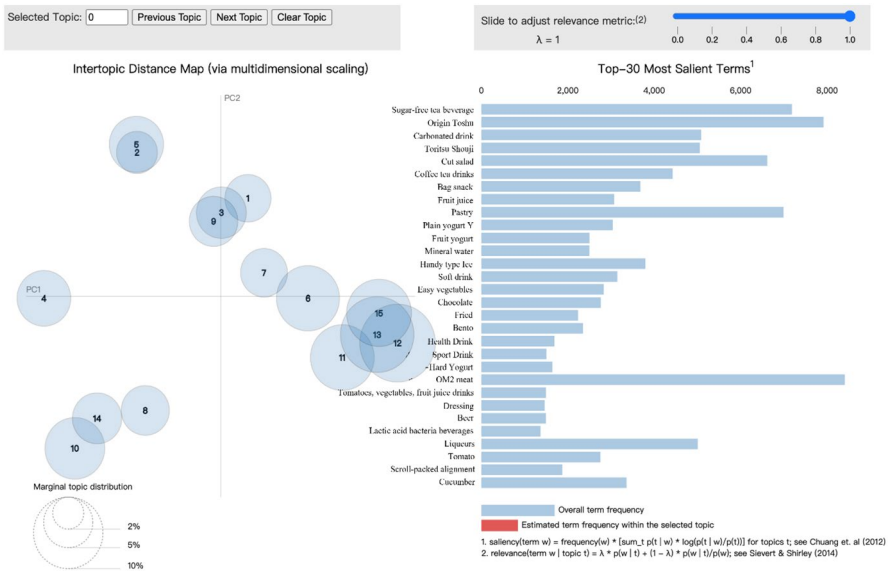


Fig. 5 The LDAvis's layout of the results of the topic model for the first half of Sep 2020

Simultaneously, we notice that the proposed framework also captures the emergence of some seasonal product categories as we expected. Some of these products appear in multiple topics in specific periods. The most direct results will be in the product category of Christmas goods, which only appears in December's topic. This result verified some of the effectiveness of this framework because we can assume that this is the only time customers would be interested in Christmas goods. Using the proposed framework, retailers of the supermarket not only found a new actionable plan to promote sales of products on the same topic, but they could also change their strategy over time in a real-time manner.

5.2 Visualization Design Based on the Results of Topic Model

After obtaining the topic model results, we decided it was equally important to represent these results better. As we explained before, simply handing these results to supermarkets in text-based form might make it difficult for retailers to make decisions about targeting promotions, especially when the number of topics tends to be high (while the number of topics in this result is somewhat tightened, in future practice, this number is also likely to be greater than 40).

Following the development of prior studies [28] in this area and discussions with marketers, in this research, we designed a three-layer visualization approach. The first layer reveals the relations among different topics using Sievert and Shirley's method called LDAvis, which is demonstrated in Fig. 5. The visualization they proposed has two panels. The left one plots the topic as a circle whose centers are determined by computing the distance between topics and using multidimensional

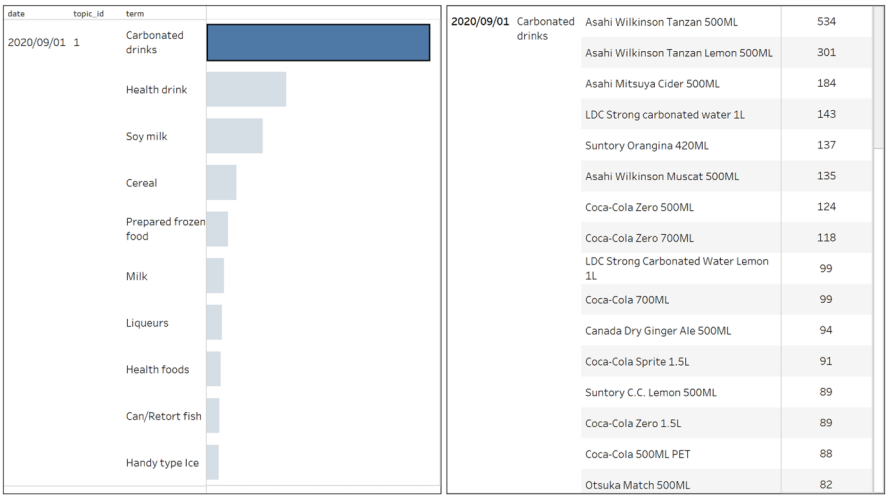


Fig. 6 Product categories in a topic at the first half of September 2020 and products in the highlighted category: Carbonated drink

scaling to project the inter-topic distances onto two dimensions, as in Chuang's research [29]. The axis in this panel does not have any specific meaning. Instead, it demonstrates the relations among topics by calculating their distances through the Jensen–Shannon divergence, and the size of the circle represents the topic distributions within the documents. The right panel is a bar chart showing the top 30 most valuable product categories for all topics on the left. When a topic is selected in the left panel, the right panel will also highlight the top 10 product categories that have higher probabilities of appearing in that topic. λ is the parameter that determines the weight given to the probability of product categories under topics. Prior studies use its value to adjust the balance between the word's probability and its lift value to obtain a more meaningful result [28], but in this research, we default the parameter to 1.

The position of circles in Figure 5 is divided into several segments [1]. For example, when we select 1, 3, and 9 in the left panel, we discover that most of the product categories in these topics are beverage-related, while 12, 13, 15, and 11 are more related to agricultural products, perhaps representing the tendency of some customers who cook their food. Accordingly, we can conclude that when the topic model is run with supermarket data, it is more valuable to visualize their relevance than for topic models on text-based data.

The second layer of visualization displays the correlation between product categories in the topics and the actual product in such product categories. Figure 6 shows the topic results (topic ID = 1) in the first half of September 2020.

[1] "Origin Toshu", "Toritsu Shouji" and the following "OM2" in Fig. 5 are the brand names of Japanese companies.

The table on the left is aligned with the product category in the topic, and while the panel on the right shows the products in that category we highlighted that they were purchased during that period. Even though we use POS data at the product category level to capture customer's buying tendencies through the topic model, the categories alone are insufficient when retailers consider marketing strategies. As mentioned earlier, some product categories contain hundreds of SKUs. For example, the highlighted "Carbonated drink" has 54 SKUs purchased in the first half of September 2020. When targeting promotions for this category, it is hard to imagine promoting 54 products simultaneously. Of these, only some products would be selected.

Finally, we observe that the combination among product categories changes over time. The basic idea of this visualization is to consider the links between categories within the same topic as a "network." This network may not change, but depending on the characteristics of supermarket products, the products connected to it in a network can change dramatically from one period to another for some specific product categories. Here, let us take the example of the product category "dipping source" in Table 5.

Table 5 illustrates in our proposed framework (in our proposed visualization how the combination of dipping sauces with other product categories evolved over time). Dipping sauce appears in four topics from four different periods after November onwards. We assume that the first topic represents the dishes that combine Japanese and Western styles based on product categories, such as green peppers, chicken, and beefsteak. The second topic is assumed to be a seafood nabe dish because of the presence of seafood products. The third is assumed to be a traditional Japanese Kanto dish based on the category of Oden set, and the last is assumed to be a nabe dish based on bean sprouts, mushrooms, and pork. As topics disappear and emerge, the product categories associated with dips vary. It is essential to visualize this "network" between product categories to understand this variation and make more targeted marketing efforts. Moreover, Table 5 also displays the necessity of visualizing specific products within a product category in the second layer of visualization. Even for the same product category (like "dipping sauce"), those products that need to be promoted can be very different under different topics depending on the products' commodity characteristics. Furthermore, Table 6 demonstrates that such changes cannot be discovered using DTM, which only shows one pattern. This result indicates that models like DTM may not be applicable to supermarket data, because they look at continuous changes in topics. In contrast, the results of this research show that on supermarket data, changes in topics will be more discrete in nature. Topics and product categories like those in Table 5, which are only possible in the winter, may have been missed if we had not split data and run the models separately. Schmidt also addressed similar concerns about using topic models from the perspective of humanities research. In his words, "Dirichlet distributions are convenient abstractions for topic-document distributions but are an obviously incorrect prior for topic-year distributions" [30].

Table 5 Combinations of dipping sauces with other product categories evolved based on the results of the proposed framework

	Period	Product category	Period	Product category
Dipping source in the proposed framework	2020/11/16	OM2 meat	2020/12/01	Spices
		spices		Spirits
		green pepper		Well-pickled vegetables
		mushroom		Seafood dried food
		Differentiating chicken		Shrimp
		Beefsteak		Seaweed
		Beef trimmings		sugar
		ginger		Pork side
		vinegar		Soy sauce
	2020/12/16	Oden set	2021/01/01	Enoki mushrooms
		Oden's ingredients		Green onions
		Hanpen		Bean sprout
		OM2 meat		OM2 meat
		Beer		General soy tofu
		Liqueurs		Chinese Ethnic
		Distilled shochu		Chinese cabbage
		Satsuma-age		Pork trimmings
		Tofu-related		Diced chicken

Table 6 Combinations of dipping sauces with other product categories evolved based on the results of DTM

Product category	
Dipping sauces	Chopped pork meat
Easy vegetables	Ramen
Bean sprouts	Udon
Cabbage	Noodle-related ingredients
Baked products	Thinly sliced pork meat

Table 7 Comparison of the number of topics and product categories in the proposed framework, DTM and LDA (the number of topics for both DTM and LDA = 30)

Period	The number of topics for the proposed framework	The number of product categories in topics (proposed framework)	The number of product categories in topics (DTM)	The number of product categories in topics (LDA)
2020/9/1	15	114	227	236
2020/9/16	19	138	229	236
2020/10/1	13	102	226	236
2020/10/16	13	107	225	236
2020/11/1	30	217	223	236
2020/11/16	33	241	221	236
2020/12/1	33	240	222	236
2020/12/16	33	242	223	236
2021/1/1	17	129	225	236
2021/1/16	36	257	225	236

5.3 Evaluation

After implementing our proposed framework, retailers will be able to see the (1) relations between topics, (2) names and probabilities of each product category within a topic, and (3) the changes in the combination of product categories over time. Our modeling and visualization solution demonstrates that instead of focusing on model improvement, sometimes it is more important to show the model's results and consider the characteristics of the different data. To illustrate the effectiveness of our approach, we are currently working with a? supermarket to run targeted promotions based on the framework that we have proposed, . However, because this experiment is still in progress, we have evaluated the preliminary feasibility of this framework using scenario settings that capture the vision of our approach, which has also adopted by several prior studies [1, 2].

The scenario analysis estimates potential gains the proposed model may offer to a supermarket and compares this analysis to the undifferentiated target promotions. Given that one of the objectives of this study is to demonstrate that seasonal product variation can be better captured through split data, we ran the LDA

Table 8 Scenario settings for the comparison of the proposed framework and LDA (or DTM)

	The growth of profits for LDA (or DTM)	The growth of profits for the proposed model
Standard promotion	5%	3%
Target promotion	–	10%

model and DTM model once each on the entire period data to facilitate comparison, and these models are defined as "undifferentiated promotions". The results of these models are shown in Table 7. The number of topics in this LDA model is also chosen by Renyi entropy, and the number in DTM is set by hand. Also, the framework we proposed shows some seasonal product categories that both LDA and DTM missed, such as Christmas goods.

With product quantity and revenue in the data, we can check whether the number of topics coincides with the generated revenue gains and the product quantity. In the results, except for the second half of December, those values obtained mainly corresponded to each other. When taking the more significant number of topics and the smaller number for comparison, the smaller number of topics will naturally have a relatively more minor product's quantity and revenue, which leads to an inability to compare directly. Therefore, coupled with the hypothesis that retailers may target some major product categories in the topics, based on the results of the first layer of visualization, we only selected the top five topics with a relatively high percentage of topics per period for comparison with the top five topics of the LDA and DTM models.

To evaluate the performance of either both targeted marketing types or campaigns, similar to Reutter's research [1], we first defined scenario settings to determine the profit lift for those campaigns. Since the results of three models (LDA, DTM, and the proposed framework) might have common product categories, the promotion of such product categories is defined as "standard promotion." On the other hand, some product categories only appear in our proposed framework; thus, the promotion of these categories will be recognized as "target promotion."

When retailers use the results of LDA or DTM to decide which product categories should be advertised, we assume a profit lift value of 5% [1, 31]. For product categories in "target promotion," the value of the growth of quantity and revenue will be set as 10% [1] when retailers promote these categories. Nevertheless, when product categories in a standard promotion and target promotion are advertised simultaneously, the customer's attention might be more focused on the target promotion, and the growth of profits for standard promotion categories might be lower. Thus, a lower value of 3% might be more appropriate. We summarized these scenario settings in Table 8.

Figure 7 shows the results of the scenario analysis of the LDA model in terms of quantity and revenue. Most of the time, except for November and January, promotions based on our proposed framework outperform the results of the LDA model run on a complete period data, which indicates that undifferentiated promotions may be more appropriate in these months. Also, the slight differences in quantity and

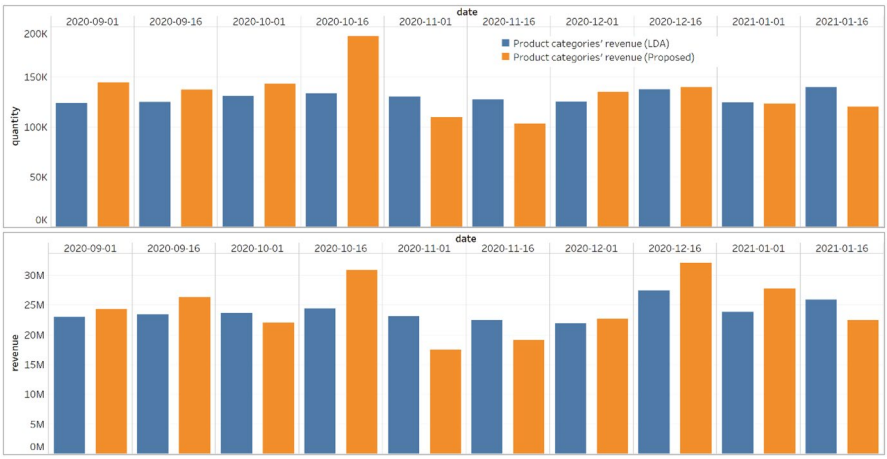


Fig. 7 The results of scenario No. 1: Compared with the LDA model by quantity and revenue

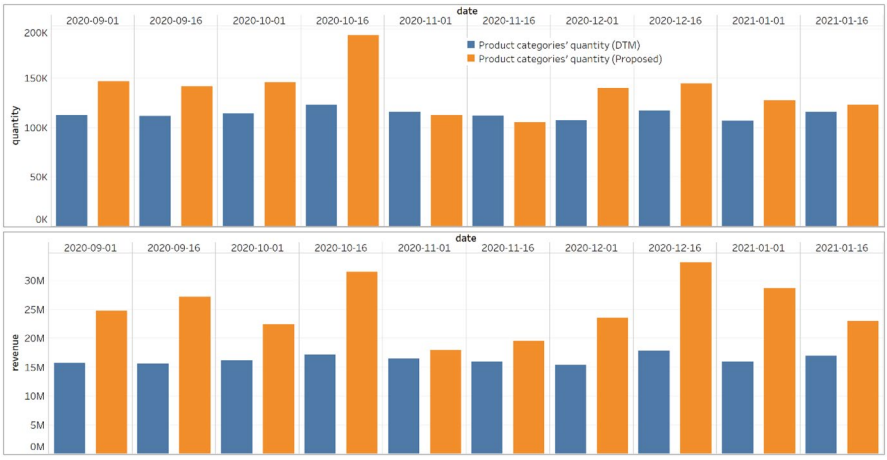


Fig. 8 The results of scenario No. 2: Compared with DTM model by quantity and revenue

revenue are reflected in the graphs. As the results for the first half of October show, higher quantity does not appear to equal higher sales, leading to different promotional strategies.

On the contrary, the results in Figure 8 illustrate that the promotions implemented according to the DTM results are lower than our framework, both in terms of quantity and revenue. The reasons for this are probably due to the distribution of topics in each period, and the combination of product categories in each topic that did not change a lot. Based on these estimation results, we can conclude that our proposed model may be a better approach for targeted marketing in supermarkets compared to the LDA model on the entire period data and the DTM model.

6 Conclusion and future research

This research proposed a framework to capture the latent patterns of customer's purchase behavior from supermarket POS data using the LDA model with Renyi entropy over five months. Instead of running the LDA model on complete period data or using the DTM model to analysis the changes of topics over time, we divided the five-month data into ten datasets and ran the LDA model separately. The estimation results show that the proposed framework might be better to observe these changes and capture specific seasonal product categories. In addition, to ensure that the topics for each period contain the most useful information, we used Renyi entropy to determine the optimal number of topics, which is the first time this method has been applied to POS data in a Japanese supermarket. As a result, the number of topics in each period is well summarized under 35 topics, representing a concentration of practical information. Once the topic model results were available, as the next component in the proposed framework, we designed a three-layer visualization approach to interpret these results to help retailers design target promotion strategies. The first layer showed the connections between the various topics and product categories. Retailers can use this to determine the priority of topics to promote. Next, we demonstrated the visualization to reveal what products are included in those product categories. As discussed in 5.2, it is unrealistic to promote every item in a product category. Finally, we proposed a third layer of visualization to show how the combination of product categories in topics changes over time. Although the product categories in each topic will also increase or decrease over time when using the DTM model, the topics themselves will also increase or decrease in practical use, which the DTM model might miss. If this third later of visualization is combined with the second layer, marketers can target specific products in different seasons, which also allows the topic model to be used more thoroughly.

The proposed framework can help retailers decide what products should be promoted using this framework twice a month. Likewise, this framework can also suggest what product categories will be combined as a whole, making a unique, themed promotion possible. Furthermore, when the proposed framework is applied, a loyalty program can be created to better engage customers and further enhance customer retention.

There are some limitations in this research that need to be further studied. First, although we propose the framework for visualization, those results are presented separately. A system that can combine all those visualizations is needed for the convenience of the retailers. Such a system that can integrate and present practical information about topic models across different domains is necessary for future research. Also, even though our estimation proved the validity of our framework, field experiments still need to be conducted, depending on the marketing strategy (whether to issue coupons or give more points), and the topic model could be applied in different ways and with different results. How to redesign the loyalty program in a supermarket based on the results of the topic model will be the next focus of the author's research. Furthermore, even though we split

data manually to observe changes over time instead of using DTM, this framework does not reflect the time continuum. Future research will focus on these limitations.

Acknowledgements This study has been supported partially by United Supermarket Holdings Co. Ltd. and JSPS Grant no.20K20482.

References

1. Reutterer, T., Hornik, K., March, N., & Gruber, K. (2016). A data mining framework for targeted category promotions. *Journal of Business Economics, 87*(3), 337–358. https://doi.org/10.1007/s11573-016-0823-7
2. Trusov, M., Ma, L., & Jamal, Z. (2016). Crumbs of the cookie: user profiling in customer-base analysis and behavioral targeting. *Marketing Science, 35*(3), 405–426. https://doi.org/10.1287/mksc.2015.0956
3. Annie, L. C. M., & Kumar, A. D. (2012). Market basket analysis for a supermarket based on frequent itemset mining. *International Journal of Computer Science Issues (IJCSI), 9*(5), 257.
4. Kaur, M., & Kang, S. (2016). Market basket analysis: identify the changing trends of market data using association rule mining. *Procedia computer science, 85*, 78–85. https://doi.org/10.1016/j.procs.2016.05.180
5. Ma, S., & Fildes, R. (2017). A retail store SKU promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research, 260*(2), 680–692. https://doi.org/10.1016/j.ejor.2016.12.032
6. Richards, T. J., Hamilton, S. F., & Yonezawa, K. (2018). Retail market power in a shopping basket model of supermarket competition. *Journal of Retailing, 94*(3), 328–342. https://doi.org/10.1016/j.jretai.2018.04.004
7. Ohsawa, Y. (2018). Graph-based entropy for detecting explanatory signs of changes in market. *The Review of Socionetwork Strategies, 12*(2), 183–203. https://doi.org/10.1007/s12626-018-0023-8
8. Cil, I. (2012). Consumption universes based supermarket layout through association rule mining and multidimensional scaling. *Expert Systems with Applications, 39*(10), 8611–8625. https://doi.org/10.1016/j.eswa.2012.01.192
9. Boros, P., Fehér, O., Lakner, Z., Niroomand, S., & Vizvári, B. (2016). Modeling supermarket re-layout from the owner's perspective. *Annals of Operations Research, 238*(1–2), 27–40. https://doi.org/10.1007/s10479-015-1986-2
10. Larson, J. S., Bradlow, E. T., & Fader, P. S. (2005). An exploratory look at supermarket shopping paths. *International Journal of research in Marketing, 22*(4), 395–414. https://doi.org/10.1016/j.ijresmar.2005.09.005
11. Yada, K. (2011). String analysis technique for shopping path in a supermarket. *Journal of Intelligent Information Systems, 36*(3), 385–402. https://doi.org/10.1007/s10844-009-0113-8
12. Liu, Y., & Yang, R. (2009). Competing loyalty programs: impact of market saturation, market share, and category expandability. *Journal of Marketing, 73*(1), 93–108. https://doi.org/10.1509/jmkg.73.1.93

13. Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology, 1*(1), 4–20. https://doi.org/10.4304/jait.1.1.4-20
14. Koltcov, S., & Ignatenko, V. (2020). Renormalization analysis of topic models. *Entropy, 22*(5), 556. https://doi.org/10.3390/e22050556
15. Koltcov, S. (2018). Application of Rényi and Tsallis entropies to topic modeling optimization. *Physica A Statistical Mechanics and its Applications, 512*, 1192–1204. https://doi.org/10.1016/j.physa.2018.08.050
16. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research, 3*, 993–1022.
17. Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine, 27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079
18. Sun, F. T., Yeh, Y. T., Mengshoel, O. J., and Griss, M. L. (2013). Latent topic analysis for predicting group purchasing behavior on the social web. In UAI Application Workshops, pp. 67–76
19. Iwata, T., Watanabe, S., Yamada, T., and Ueda, N. (2009). Topic tracking model for analyzing consumer purchase behavior. In Twenty-First International Joint Conference on Artificial Intelligence
20. Paul, M., and Girju, R. (2009). Cross-cultural analysis of blogs and forums with mixed-collection topic models. In Proceedings of the 2009 conference on empirical methods in natural language processing, pp. 1408–1417. https://doi.org/10.3115/1699648.1699687.
21. Blei, D. M., & Lafferty, J. D. (2006, June). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (pp. 113–120). doi: https://doi.org/10.1145/1143844.1143859.
22. Tang, J., Meng, Z., Nguyen, X., Mei, Q., and Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In International Conference on Machine Learning, pp. 190–198. PMLR
23. Iwata, T., & Sawada, H. (2013). Topic model for analyzing purchase data with price information. *Data Mining and Knowledge Discovery, 26*(3), 559–573. https://doi.org/10.1007/s10618-012-0281-y
24. Yang, Z., Kotov, A., Mohan, A., and Lu, S. (2015). Parametric and non-parametric user-aware sentiment topic models. In Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval (pp. 413–422). https://doi.org/10.1145/2766462.2767758.
25. Cao, D., Ji, R., Lin, D., & Li, S. (2016). Visual sentiment topic model based microblog image sentiment analysis. *Multimedia Tools and Applications, 75*(15), 8955–8968. https://doi.org/10.1007/s11042-014-2337-z
26. Reisenbichler, M., & Reutterer, T. (2019). Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics, 89*(3), 327–356. https://doi.org/10.1007/s11573-018-0915-7
27. Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval pp. 50–57. https://doi.org/10.1145/312624.312649.
28. Sievert, C., and Shirley, K. (2014, June). LDAvis: A method for visualizing and interpreting topics. In Proceedings of the workshop on interactive language learning, visualization, and interfaces, pp. 63–70. https://doi.org/10.3115/v1/w14-3110
29. Chuang, J., Ramage, D., Manning, C., and Heer, J. (2012). Interpretation and trust: designing model-driven visualizations for text analysis. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 443–452. https://doi.org/10.1145/2207676.2207738.
30. Schmidt, B. M. (2012). Words alone: dismantling topic models in the humanities. *Journal of Digital Humanities*, 2(1). 49–65. from http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/
31. Bijmolt, T. H., Van Heerde, H. J., & Pieters, R. G. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of marketing research, 42*(2), 141–156. https://doi.org/10.1509/jmkr.42.2.141.62296