2016 IEEE 16th International Conference on Data Mining Workshops

Method for Extraction of Purchase Behavior and Product Character Using Dynamic Topic Model

Mamoru Emoto
School of Engineering
The University of Tokyo
Tokyo, Japan 113-8654
Email: mamoemo.0130@gmail.com

***Abstract*—In this study, we focus on extraction of latent topic transition from POS data. POS analysis is conducted to obtain the frequent pattern of customer ' s behavior. The fundamental method for POS analysis is to conduct market basket analysis. By doing Market basket analysis, the sets of products that are often bought at the same time can be extracted. In market basket analysis, however, the effect of time series is not considered. We conducted the experiment based on two hypotheses. One is that each product has several topics. The other is that the proportion of each product on a topic changes as the period changes. To extract topics and their changes, we use Dynamic Topic Model (DTM), which is an extended model of Latent Dirichlet Allocation (LDA). Then we obtain the change of the topic-word distribution on each topic. Different topic has different characters, but it seems that there is a relationship between each topic. Therefore, we conduct correlation analysis to several items. From the result of visualization of product features vector of several items, we can obtain that each product has unique time-series change of product feature. This study is also conducted to reveal product features vector based on Customer Purchase Behavior. By using DTM, each basket is transformed into probability distribution vector, and we use this topic vectors as each basket's evaluation result of topic features. We divide POS data into 12 groups by purchasing time and create a heat map that indicates changes of topic proportion as time advances. By conducting this analysis, we can grasp customer behavior based on the topic vector space. These analysises reveals product features are created based on topic correlations and its change and customer behavior can be extracted as the change of topic proportion, so the results show that the presented method is promising in the extraction of products' features and customer behavior.**

I. Introduction

It is important to analyze purchase data collected from POS(Point Of Sales) system. POS analysis is conducted to obtain popular products, a difference of purchased products according to each area, and the combination of goods which can frequently be bought at the same time. This analysis is conducted for deciding the whole design of the shop and strategy of sales events. In a sense, the result of the POS analysis is useful as a bridge that connects shop staffs and customers. An increasing number of companies collect ID-POS that link ID number of customers to POS. Applying machine learning to ID-POS enables us to understand each customer's preference and recommend products according to it.

One of the traditional POS analysis methods is market basket analysis. This analysis is a kind of association analysis, which enables us to extract a combination of products which are frequently bought at the same time. Association analysis uses a set of transactions to discover rules that indicate the likely occurrence of an item based on the occurrences of other items in the transaction. There are three association rule evaluation metrics:

- Support
- Confidence
- Lift

Lift value implies the validity of the rule. In the context of the basket analysis, the rule means the tendency that certain products are bought in the same transaction. For example, the following rule can be extracted from pos data:

$$\{Diapers\} \rightarrow \{Beer\}$$

If the lift value of the rule is high, it suggests that a strong relationship exists between the sale of diapers and beer because many customers who buy diapers also buy beer at the same time. This kind of rules is useful for retailers to help them identify new opportunities for cross-selling their products to the customers.

However, association analysis does not take the effect of time-series into consideration. Besides, at the real market where customers buy products, it can be assumed that each product has several topics(= characters) and its likelihood of each topic changes as time advances. For example, in Japan, fried chickens are sold at delicatessen section of supermarket stores. There are two kinds of customers who buy them. One is customers who buy them as one of the dishes for lunch. The other is who buy them as midnight snacks for the fun of drinking beer. As we can recognize from the example above, even if customers buy same products, each products' likelihood of topics differs according to the purchasing time. Therefore, taking the change of topic-likelihood into consideration enables us to extract further preferences of

2375-9259/16 $31.00 © 2016 IEEE
DOI 10.1109/ICDMW.2016.12

customers that have not be obtained from simple association analysis, which is useful to decide marketing strategies based on the result of POS data. In this study, we apply Dynamic Topic Model(DTM)[1], one of the topic models mainly used in the field of natural language processing. DTM is the extended model of Latent Dirichlet Allocation(LDA)[2]. LDA enables us to obtain the topic distribution of the document. DTM takes time-series change of the topic characters into consideration. We assume that this model appropriately fits the real customer behavior and apply it to POS data to extract topics of customers behavior and their change. We also create product vectors based on the result of DTM.

II. Related works

Authors[3] conducted the study of the method to extract features of each recipe and classify recipes on recipe-sharing websites. On recipe-sharing websites, There are many recipes and reviews posted by users. Every recipe and review data has the date when they are posted. For the review data, morphological analysis is conducted. By using word counting method, four-dimensional vectors are obtained. The factors of the vector are 'Health', 'Family', 'Easy-cooking', 'Taste'. These factors are defined from Action Planning(AP)[4], the workshop method for making innovative ideas more practical. The study reveals that particular months has high value at some factors compared with other months.

Takahashi et al.[5] proposed a method to identify bursts of topics. The proposed method combine the burst analysis and DTM that is used in this study to extract relevant information from time-series news data. This method focuses on the burst of each topic, and the input data is news text data.

Ishigaki et al.[6] applied topic model to POS data to classify customers and products based on the customers' purchase behavior. Probabilistic Latent Semantic Indexing(PLSI), one of popular topic models, is used to extract purchase topics from ID-POS data. PLSI is a kind of clustering method so by applying it to ID-POS data, customers and products can be classified at the same time. Then additional data like weather, date, time and the effect of a sale are linked to ID-POS data. Then relationships between each variable of combined data are expressed as probabilistic structure model by using Bayesian network. The other study [7] focuses on the classification of customers and products based on the purchase behavior and personality obtained from a questionnaire survey. Like the analysis flow above, at first, they apply PLSI to ID-POS data to classify customers and products at the same time. Then, factors of daily living and purchase behavior are extracted by factor analysis and combined with the classification result of PLSI. The purpose of this approach is the classification of customers and products with each customers' preference and behavior. Both studies focus on extracting customers' characters and clustering of customers by using topic model, and time-series topic change of a target store is not taken into consideration. In this study, we assume topic characters of a target store do time-series change. Therefore, DTM analysis is conducted to obtain topic transition of the target store. ID-POS analysis and creation of product feature vector are also conducted based on the result of DTM analysis. This approach is noble one for market basket analysis. By applying topic model, the transition of the characters of basket and product can be discussed at the topic vector space. We focus on this feature to evaluate and analyze the transition of purchase behavior, which will bring about the further understanding of customers for retailers.

III. Dynamic Topic Model

Dynamic Topic Model(DTM) is an extended model of LDA. A lot of extended Topic models are proposed. Joint Topic Model[8] and Correspondence Topic Model[9] are the models that introduce the idea of side information to Topic Model. By applying these models, we can analyze document data with side information like review points. Correlated Topic Model[10] is proposed to consider and model the correlations of topics. In this study, DTM is applied to POS data. This model proposes to apply Markov property to document-topic distribution and topic-word distribution, which enables us to obtain time-series changes of topic proportion and topic-word assignment. In a original topic model, each document is assumed drawn from the following generative process:

1) Choose topic proportions θ from a distribution over the $(K-1)$-simplex like a Dirichlet.
2) For each word:
   a) Choose a topic assignment $Z \sim Multinomial(\theta)$.
   b) Choose a word $W \sim Multinomial(\beta_z)$.

This process implicitly assumes that the documents are drawn exchangeably from the same set of topics and that topics are fixed and do not evolve. However for many collections, the order of the documents reflects an evolving set of topics. In DTM, analyses are conducted on the presupposition that the data is divided by time slice with a $K-$component topic model, where the topics associated with slice t evolve from the topics associated with slice t-1. The graphical model is often used to express Probabilistic Generative Models. Fig. 1 shows the graphical model of DTM. Each topic's natural parameters β evolve over time, together with the mean parameters α_t of the logistic normal distribution for the topic proportions. The generative process of DTM is as follows:

for t = 1,2,...,T:
  1.draw topics
  $\beta_{t,k} \mid \beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma^2 \boldsymbol{I})$
  2.draw α_t
  $\alpha_t \mid \alpha_t - 1 \sim N(\alpha_{t-1}, \delta^2 \boldsymbol{I})$
  for each document:
    3.topic proportion
    $\eta_{t,d} \mid \alpha_t \sim N(\alpha_t, a^2 \boldsymbol{I})$
    $\theta_{t,d} \mid \eta_{t,d} \sim \pi(\mid \eta_{t,d})$
    for each word:
      4.topic-word assignment
      $z_{t,d,n} \mid \theta_{t,d} \sim Multinomial(\theta_{t,d})$
      5.word observation
      $w_{d,n} \mid z_{d,n}\beta_{t,k} \sim Multinomial(\pi(\beta_{t,z_{d,n}}))$

Here, K, N, D mean the number of topics, vocabularies, and documents. α is a hyperparameter for generating document-

topic proportions. α_t is generated based on α_{t-1}. To create time-series topic model, state space model is applied to LDA, which is the basic expansion of DTM. On state space model, a normal distribution is assumed to be used. Therefore, a real vector is generated by a model based on state space model. Note that soft-max function π maps the multinomial natural parameters to the mean parameters, $\pi(\boldsymbol{v}) = \frac{\exp(v_k)}{\sum_l \exp(v_l)}$. η is a real vector to generate document-topic proportion and θ is document-topic proportion made from η that is transformed by the soft-max function π. β is real vector generated as topics, and the Markov property of β means the topic-word proportion of each topic changes according to time slices. topic-word proportion ϕ is generated from β transformed by the soft-max function.

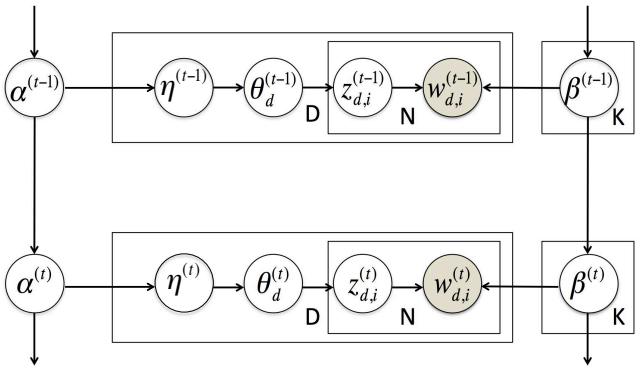

Fig. 1. The Graphical Model of DTM

IV. Extraction of Purchase Topic by Dynamic Topic Model

A. Analysis Flow and Experimental Condition

In this study, we applied DTM to POS data to extract purchase behaviors as changes of topic proportions. The information of POS data is shown in I and the analysis flow is as follows:

- Make basket data based on transaction ID of POS data
- Make input data by extracting baskets filtered by the number of purchase items(minimum number:5)
- Run DTM analysis
- Obtain output and extract changes of topic-word proportion

TABLE I
Collected POS

Period	2014/09/01-2014/11/30
Number of filtered baskets	212439
Number of vocabularies(= products)	17162

In many cases, Topic model is applied to text(document) data and the input type is Bag of Words. In this study, we used it to POS data because there is a structural analogy between documents and purchase baskets. That is, both documents and purchase baskets can be described as Bag of Words. We used C++ program to run DTM and the initial condition of DTM was decided according to preliminary experiments (topic number k = 5, α = 0.1, time period = day (t = 0,1,2, ... , 90)).

B. Result:Product DNA analysis

Table II shows each topic's character. Rough characters of topics are fixed but the topic-word distribution changes as time advances. This character reflects each topic's trends in which there are seasonal changes about products taken by customers.Fig.2 and Fig.3 show top4 products of each topic.

TABLE II
Topic Features

Topic	Feature
Topic1	side dish, salad, ready-cooked foods
Topic2	ingredients(vegetables, meat)
Topic3	sweetened buns, drinks, snacks
Topic4	side dish, soda, alcohol
Topic5	dairy products, meat

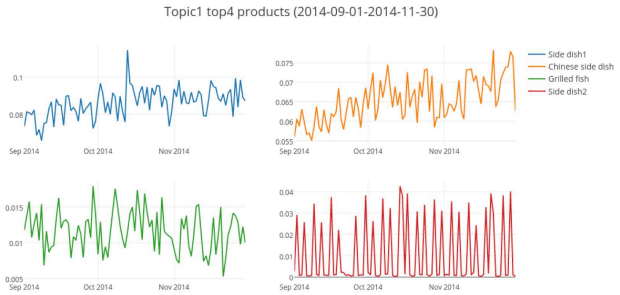

Fig. 2. Topic1: Top4 products

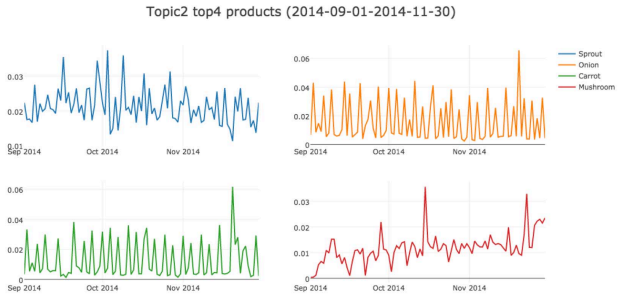

Fig. 3. Topic2: Top4 products

This result tells us each product's significance in each topic. We extract these four products based on the mean value of a topic-word distribution because if the value of the topic-word distribution of a product in a topic is high, it means the influence of the product is significant in the topic. From Fig. 2, we can recognize that side dish1 and Chinese side dish

constantly indicate high value, which means that in topic1, the influence of side dish products is significant. On the other hand, Fig.3 shows vegetables are dominant in topic2. Every topic has a different feature, and it means that by using DTM, we can analyze and discuss basket data and product features based on the topic vector space. Besides, focusing on the each product's topic distribution itself, we can grasp products' features.

Fig.4 and Fig.5 show changes of topic-word distribution about two products(Squid sashimi and side dish). Here x-axis means period and y-axis means each topic's value of the topic distribution.Fig.4 indicates that there are two correlations. One is between topic 1 and topic 5, and the correlation coefficient is -0.7316. The other is between topic 2 and topic 4, and the correlation coefficient is 0.4581. On the other hand, based on Fig.5 there is a correlation between topic 1 and topic 5 and the correlation coefficient is -0.3690. Considering analysis results, it is presumed that there are some correlations between each topic and that some products have unique correlations on the change of Topic-Word distribution. This result can be applied to create new market strategies that are not based on conventional methods like market basket analysis. For example, this result enables us to classify each product according to the similarity of the topic change and this classification is based on the time-series topic change. In real market, popular products change according to seasons. Therefore, if some products indicate high similarity on change of Topic-Word distribution, this means that these products have similar character in terms of topic, which may enable us to decide the appropriate merchandise arrangement and special offers.

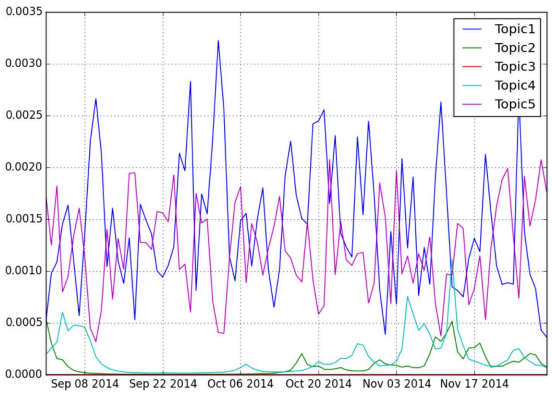

Fig. 4. Change of Topic-Word distribution: Squid sashimi

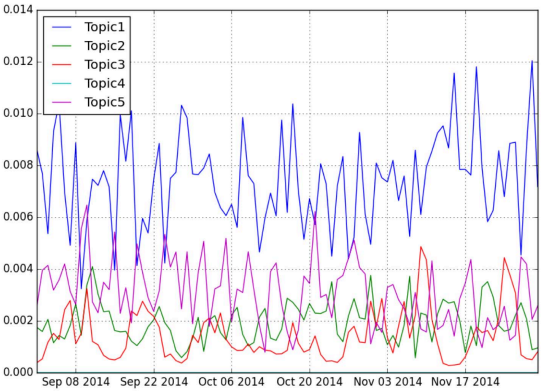

Fig. 5. Change of Topic-Word distribution: Side dish

C. Result:Customer Behavior analysis

Fig.6 shows the change of topic proportion of baskets on a weekday and Fig.7 indicates the difference of topic proportion change between weekend and weekday. DTM transforms each purchasing basket into a topic vector. The topic vector is probability distribution and factors mean each basket's likelihood of topics. As shown in Fig.3, topic2 is the ingredient topic. Therefore for example, if a customer buys four ingredients and one snack in one transaction, the topic vector of this purchase basket indicate a high value on topic2. This transformation enables us to analyze each basket's purchase feature based on the topic proportion. We assume that there are time-dependency in each topic and changes of the topic proportion of baskets in a day. Hence we divide baskets into 12 groups according to the purchasing time and calculate the mean value of topic vectors in each group and visualize as the heat map. From Fig.6 and Fig.7, we can recognize that the topic proportion changes according to time and each topic has a tendency to indicate high value at a certain time. Topic2 indicates high value in morning and afternoon but at night the value becomes lower. Instead, topic1 indicates high value at night. Topic2 is ingredient topic, and the value is around 0.3. This result deserves more than a passing notice. According to the interviews to supermarket staffs, the result corresponds to the aim that the sales of ingredients occupy 30 percent of sales amount. Focusing on topic5, compared with the weekend purchase pattern, more people do this kind of purchase on a weekday. Moreover topic 1 and topic 4 shown in Fig.6 are both characterized by side dish but changes of topic proportion differ and considering this, we can recognize that each product has some topic characters, and this character changes as time advances. By conducting this analysis, we can grasp customer behavior based on the topic vector space and its change and these result can be applied to decide marketing strategy of sales events and understanding each store's character.

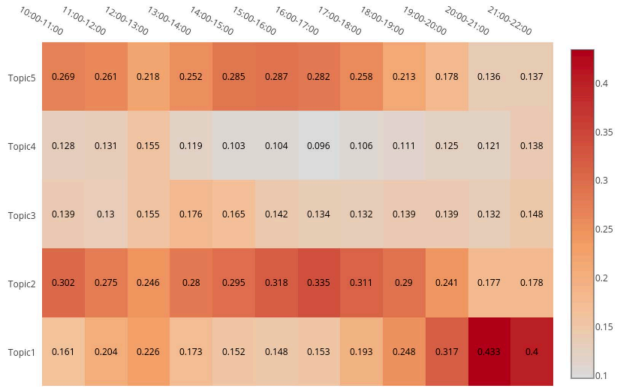

Fig. 6. Change of Topic proportion of baskets(Weekday)

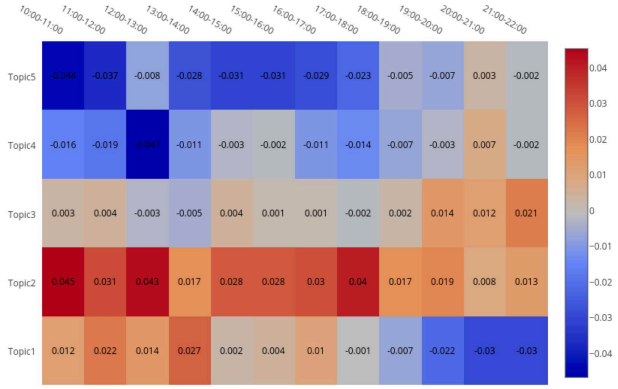

Fig. 7. Change of Topic proportion of baskets(Weekend-Weekday)

V. Conclusion

In this study, a method for extracting products' features and customer behavior is proposed. We apply Dynamic Topic Model(DTM) to POS data collected from September 1, 2014, to November 30, 2014, and extract product features and customer's purchase behavior. By applying the concept of the topic to POS analysis, we can discuss both customer behavior and product features based on the same metric: topic vector space. That is, this enables us to create all product vectors and purchase baskets as k-dimension topic vectors (k= number of topic). We reveal that topics created from DTM analysis have different features, and there are some correlations between topics in each product. We also find out that the topic proportion of purchase baskets changes as time advances in one day because customer purchase patterns tend to change and that this result corresponds to market strategies and knowledge from experience of professionals about real purchase behavior. This analysis is promising for creating and improving marketing strategies based on real data. In the future work, we will extend the period and extract long-term topic changes. We will also conduct further research about correlations of topics. This research will lead us to new clustering based on similarities of topic changes. Visualization method should be improved. In this research, we just create time-series plots. We can obtain topic vectors from DTM so we will take advantage of this feature for future analysis and visualization. The evaluation of this method should also be considered. This analysis is conducted to find business opportunities and hidden patterns so we will carry out some experiments at real supermarket stores to evaluate the analysis result.

Acknowledgment

This study has been supported by JST CREST and JSPS KAKENHI Grant Numbers JP16H01836, JP16K12428. POS data are supplied by KASUMI CO., LTD. I would like to express my gratitude for their support.

References

[1] Blei, D.M. and Lafferty, J.D., "Dynamic Topic Models," Proc. 23rd ICML, pp.113-120, 2006.
[2] Blei, D.M., Ng, A.Y. and Jordan, M.I., " Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol.3, pp.993-1022, 2003.
[3] Emoto, M. and Ohsawa, Y., "Feature Extraction and Recipe Classification based on Recipe-Sharing Websites," IEICE Technical Report, Vol.115, No.337, pp.73-77, 2015.
[4] Hayashi, T. and Ohsawa, Y., "Processing Combinatorial Thinking: Innovators Marketplace as Role-based Game Plus Action Planning," International Journal of Knowledge and Systems Science, Vol.4(3), pp.14-38, 2013.
[5] Takahashi, Y., Yokomoto, D., Utsuro, T. and Yoshioka, M., "Analyzing Burst of Topics in News Stream," IPSJ SIG Technical Report, Vol.2011-NL-204, No.6, pp.1-6, 2011.
[6] Ishigaki, T., Takenaka, T. and Motomura, Y., "Method for automatic extraction of relations of context-dependent variables of each category from ID-POS data", The Operations Research Society of Japan, Vol.56, No.2, pp.77-83, 2011. (Japanese)
[7] Ishigaki, T., Takenaka, T. and Motomura, Y., "Customer and Item Categorization by Integration of ID-POS Data and Questionnaire Data by Using Probabilistic Latent Semantics Indexing", IEICE Technical Report, Vol. 109, No.461, pp.425-430, 2010.
[8] Flaherty, P., Giaever, G., Kumm, J., Jordan, M.I., and Arkin, A.P., "A latent variable model for chemogenomic profiling", Bioinformatics, 21(15), pp.3286-3293, 2005.
[9] Blei, D.M. and Jordan, M.I., "Modeling annotated data", In Proceedings of Annual ACM SIGIR Conference, SIGIR, pp.127-134, 2003.
[10] Blei, D.M. and Lafferty, J.D., "A correlated topic model of science", The Annals of Applied Statistics, pp.17-35, 2007.