IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 3, 2020, accepted September 18, 2020, date of publication September 28, 2020,
date of current version October 9, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3027023

Similarity Measure for Product Attribute Estimation

PATRICIA ORTAL[1], (Member, IEEE), AND MASATO EDAHIRO[2], (Member, IEEE)
[1]Rakuten Institute of Technology, Rakuten, Inc., Tokyo 158-0094, Japan
[2]Graduate School of Information Science, Nagoya University, Nagoya 464-8601, Japan

Corresponding author: Patricia Ortal (patricia.ortal@ieee.org)

ABSTRACT Representing products as a combination of properties that capture the essence of consumer sentiment is critical for companies that strive to understand consumer behavior. A catalogue of products described in terms of their attributes could offer companies a wide range of benefits; from improving existing products or developing new ones, to improving the quality of site search and offering better item recommendations to users. In this paper, we propose a method that encodes products as a sequence of attributes, each of which represents a different dimension of the consumer perception. In the proposed method, first, a base product set with known attribute values is built based on consumers' perceptions. Then, new product attribute vectors are estimated using product similarity. The proposed method also incorporates a new similarity measure that is based on purchase behavior and which is suitable for estimating product attribute vector distances. Because it takes into account the magnitude of the individual components of the vectors under comparison, the proposed method is free from the limitations of conventional similarity measures. The results of experiments conducted using real-world data indicate that the proposed method has superior performance compared to conventional approaches in terms of mean absolute error (MAE) and root mean squared error (RMSE).

INDEX TERMS Attribute estimation, collaborative filtering, consumer behavior, e-commerce, similarity measures.

I. INTRODUCTION

Comprehensive market analysis and deep understanding of end users is essential to create valuable market positions and stay ahead of competitors. Identifying the consumer subjective perceptions, motivations, and preferences helps companies improve existing products so they can be customized accordingly. In-depth knowledge of consumer sentiment could also be exploited to uncover unmet product needs and generate new product development opportunities.

Representing products in terms of attributes that denote consumer perceptions is also crucial for improving the quality of e-commerce site search. Attributes are the building blocks of the product catalogues that allow e-commerce retailers to identify, organize, standardize, and display information to users [1]. They are implemented to build more efficient product recommendation systems [2], [3], filter search results more effectively [4], and ultimately improve product discoverability, thereby positively influencing sales [5]. Therefore, it is of great interest to devise methodologies that could accurately estimate product attributes.

An example of a field that addresses questions of these kinds is text mining. Text mining is an active area of research that focuses on the development of statistical techniques and machine learning algorithms that extract meaningful information from unstructured or semi-structured text. Attribute extraction involves generating attribute-value pairs from text available in product descriptions or reviews. However, these methods rely heavily on large quantities of text data [6]–[8], and are inadequate for inferring the attributes of products that are newly introduced to the market.

In conjunction with text mining, surveys constitute a common source of information on the perceptions and attitudes of a population towards certain products. For decades, companies that carry out qualitative research have been relying on surveys to understand the factors that influence buying behavior. However, conducting surveys becomes more expensive and time-consuming as the number of products and attributes increases. For online retail companies, which often have a large inventory of products, it becomes unfeasible to build an

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

attribute product catalogue based on surveys exclusively. For this reason, it is of great interest to propose methodologies that can accurately estimate product attributes without the need for large sources of user generated text or annotated data.

The main contribution of this study is a method for encoding products as a sequence of attributes, each of which represents a different dimension of the consumer perception. In the proposed method, user perceptions on a predefined product base are first obtained (via surveys) and analyzed. Then, the attribute vectors of different products from the same group are estimated using product similarity. In particular, one of the main objectives of this research was to propose a new similarity measure that is suitable for estimating product attribute vector distances based on purchase behavior.

The practical value of our proposed method is best appreciated in marketing and e-commerce applications where resources are limited and acquiring additional data can be time-consuming or expensive. Rather than increasing the survey size, our proposal generalizes from already available data. In this manner, marketers can use the results of their existing product sentiment surveys to estimate the attributes of similar products that were not originally surveyed. By building a more exhaustive product attribute catalogue in this manner, marketers may, for example, obtain advanced insights through analyses of sales trends by specific product attributes. Our proposed approach can also be used to complete missing attribute information in e-commerce product catalogues. Cleaner and more comprehensive catalogues can increase customer satisfaction through a better shopping experience with superior search results and more accurate product descriptions.

II. RELATED WORK

A. ATTRIBUTES IDENTIFICATION

Categorization allows us to grasp the maximum amount of useful information with the least cognitive effort [9]. Representing objects in terms of their attributes helps us to simplify our perception of the world in an efficient manner. We can filter out useless information and quickly identify objects with respect to their differences and similarities.

Tagging is a form of attribute assignment, since tags are used to describe particular characteristics of objects. Whereas annotating digital content has been possible for years [10], it only became popular when the weblogging community needed new ways to organize their information for easier recall and discovery. This form of tagging, where users explicitly add keywords in the form of metadata to shared content, is commonly known as collaborative tagging [11]. Some of the limitations of this classification system include the use of ambiguous terms, imprecision, and a lack of consistency, given the absence of a reference hierarchical structure [12]–[14].

Product attribute vectors can also be built automatically. This area of research falls within the domain of information extraction, which focuses on developing different approaches to automatically extract information from unstructured or semi-structured text. Some of the work in this field include different techniques to find values of predefined target attributes that describe products' intrinsic properties, such as brand name, color, or size [1], [3], [4], [15], [16]. The sources of these attributes are commonly product profiles, titles, descriptions, or reviews. An advantage of these techniques is that product attributes do not need to be defined in advance, hence, new concepts can be discovered or the most popular characteristics can be identified. However, these characteristics rarely contain user opinions [17].

Sentiment analysis, on the other hand, focuses on analyzing text to identify the affective state of a user towards specific products or services. Aspect extraction is a subtask aimed at the identification of aspects (attributes) and their associated sentiments. State-of-the-art research in this field investigates ways to efficiently extract aspects, identify the associated opinions, determine the polarity of the opinions, and do so in multiple granularities. However, owing to the high complexity of this procedure, there is little work combining all these tasks at once [18]. Moreover, existing algorithms are not adequately accurate; they are usually limited to specific segments, such as electronic products or hotels reviews [19].

In the context of consumer behavior analysis, both collaborative and automatic tagging offer a powerful way to understand the general perception and sentiment of end users towards particular products. However, these approaches are not sufficient when the objective of the research is to investigate specific product attributes, because the attributes might never appear as tags or as parts of reviews. Moreover, representing product attributes as a collection of tuples of attribute-value pairs, is not informative of the strength (intensity) with which each attribute describes each element, but only whether the property exists or not. Finally, because these methods rely on tags or text data available online, if there are no reviews available for a product, no attributes can be extracted.

The Tag Genome in [20] introduced the concept of "item genome" to encode movies' most relevant attributes. Similar to how organisms are described by a sequence of genes, the item genome encodes items in an information space based on its relation to a common set of attributes. The Tag Genome G is defined as a collection of relevance values for all tag-item pairs in $T \times I$, represented as a tag-item matrix, where T is a set of tags and I is a set of items. The relevance is calculated using a regression model on predefined features, and has been shown to have high prediction performance. However, this approach relies on training the model on thousands of labeled examples extracted from expert-maintained data or collaborative tags and text sources such as reviews and blogs.

Similar to the Tag Genome, there are a few other works in the literature that approach the prediction of tag data using different machine learning techniques. Examples of these include the prediction of tags in music using AdaBoost [21], the use of nearest neighbor classifiers to predict video features [22], and the tagging of web pages through association rule mining [23]. However, all of them rely on considerable amounts of crowd-sourced data.

The present research seeks to encode product attribute values using a matrix representation similar to the one in [20]. However, we undertook a different approach for calculating such values. Our data came exclusively from a survey, and the attributes were fixed by domain experts. Because surveys constitute a costly data-collection method, our data source was very limited. To train machine learning models with acceptable performance, a considerable number of samples were required, hence, this approach was discarded. Instead, inspired by the assumptions in the collaborative filtering technique, we first calculated the similarity between products, and subsequently computed the new product attribute vector as a weighted average of the base products, with the weights corresponding to the similarity between products.

B. SIMILARITY MEASURES

Similarity measures play a critical role in various disciplines; from biology and sociology to data analysis and machine learning. The choice of an effective similarity measure depends on the nature of the data [24]. For this reason, numerous measures have been proposed as a result of research efforts in several domains.

In general, similarity measures quantify the degree of similarity between two objects. They are particularly important in the context of collaborative filtering for recommendation systems, because they serve as a set of criteria to select groups of similar users, whose preferences are aggregated and exploited as a basis to infer other users' tastes. Consequently, the similarity computation has a direct and significant influence on the performance of collaborative filtering methods [25]. Because the underlying assumption of our proposal is similar to that of the collaborative filtering, we decided to refine our comparative analysis to the most extensively used similarity measures in this domain. Pearson's correlation coefficient (PCC), cosine similarity (COS), and Jaccard index (JAC) are among the most widely applied similarity measures to collaborative filtering tasks [26].

The similarity index is often calculated between vectors of values. For instance, in collaborative filtering, each vector corresponds to a user, and the elements of the vector correspond to ratings given to different items. Pearson's correlation measures the strength of the linear relationship between two vectors. Given two vectors u and v of rated items, PCC(u, v) is defined as follows:

$$\mathrm{PCC}(u,v)=\frac{\sum\limits_{i\in I_u\cap I_v}(r_{u,i}-\bar{r}_u)(r_{v,i}-\bar{r}_v)}{\sqrt{\sum\limits_{i\in I_u\cap I_v}(r_{u,i}-\bar{r}_u)^2}\sqrt{\sum\limits_{i\in I_u\cap I_v}(r_{v,i}-\bar{r}_v)^2}},\quad(1)$$

where I_u and I_v are the sets of items rated by users u and v, $r_{u,i}$ is the rating of item i by user u, and $\bar{r}_u$ is the average rating of user u. COS calculates the similarity of two vectors measuring the cosine of the angle between them. Orthogonal vectors are completely dissimilar, and parallel vectors are maximally similar. Given two vectors u and v, the cosine similarity is calculated as follows:

$$\mathrm{COS}(u,v)=\frac{\sum\limits_{i\in I_u\cap I_v} r_{u,i}r_{v,i}}{\sqrt{\sum\limits_{i\in I_u\cap I_v} r_{u,i}^2}\sqrt{\sum\limits_{i\in I_u\cap I_v} r_{v,i}^2}},\quad(2)$$

where I_u, I_v, and $r_{u,i}$ have the same meaning as in (1). JAC measures the similarities between two vectors calculating its overlap. Given two vectors u and v, it is defined as follows:

$$\mathrm{JAC}(u,v)=\frac{|I_u\cap I_v|}{|I_u|+|I_v|-|I_u\cap I_v|},\quad(3)$$

where I_u and I_v are the set of items rated by user u and v respectively.

Over the years, researchers have conducted comprehensive surveys [27]–[29], and have compared the advantages and disadvantages of these measures, such as in [30]–[32], and [33]. Most drawbacks come from PCC and COS, which consider exclusively the direction of the rating vectors and ignore their length; this can lead to misleading similarity scores. JAC does not consider the actual values of the ratings, only whether the item was rated or not.

In a different line of research, semantic similarity metrics have been successfully used to calculate the distance between items based on their high-level descriptions. For example [34], [35], and [36] discuss the use of different semantic similarity measures to calculate the resemblance of proteins based on their function, given by their ontology. The present work, however, focuses on the similarity measures discussed previously. That is, the similarity is to be calculated based on the explicit or implicit behavior of the users. Considering the high-level meaning of the item attributes themselves is beyond the scope of this research.

III. PROBLEM STATEMENT

We propose a method to infer the attribute values of a list of target products from a set of base products with known attribute values obtained through a survey. The attributes were selected by marketing experts with the intent to explain the most representative aspects that characterize the consumer sentiment towards a product. These attributes have an inherently low overlap, since the experts were tasked with choosing different key features related to consumer perception. For this reason, and to simplify our analysis, we assumed them to be independent.

The survey results were normalized and aggregated into a matrix with a column for each attribute $a \in A$ and a row for each product $p \in P$. The elements of the matrix are numbers between 0 and 1, indicating how relevant each attribute is to each product. A value of 0 implies no relevance, and a value of 1 indicates the highest relevance. An example is shown in Table 1. The second row shows the attribute vector for product p_2 ($[0.4, 0.8, \ldots, 0.2]$), which is hence characterized by attribute 1 with 0.4 relevance, attribute 2 with 0.8 relevance, etc.

TABLE 1. ***N* × *M* product-attributes matrix. The elements of the matrix, numbers between 0 and 1, represent how relevant to each product each attribute is.**

$$
\begin{array}{c c}
 & \begin{array}{cccc} a_1 & a_2 & \ldots & a_M \end{array} \\
\begin{array}{c} p_1 \\ p_2 \\ \vdots \\ p_N \end{array} & \begin{bmatrix} 0.3 & 0.4 & \ldots & 0.1 \\ 0.4 & 0.8 & \ldots & 0.2 \\ \vdots & \vdots & \vdots & \vdots \\ 0.7 & 0.2 & \ldots & 0.4 \end{bmatrix}
\end{array}
$$

Our task is to estimate the attribute vectors of products that have not been surveyed but belong to the categories for which data are available. Our proposal focuses on the users who purchased the products, and applies the basic idea of collaborative filters [26]. In other words, users who chose what to buy based on their set preferences. In such a case, a product purchased by a given user is more likely to have attributes in common with other products purchased by the same user. This allows us to compute similarity scores between products, and infer unobserved values from the available ones.

In summary, the underlying assumptions of our proposal are as follows:

- Consumers buy products with attributes that are consistent with their tastes.
- Similar products have similar attributes.

We consider the frequency of purchase as an indicator for preference towards a product. Therefore, we examine the total number of times each item was bought by each user within a fixed period of time. These are the data we use to calculate product similarities.

The conventional similarity measures applied in collaborative filtering have inherent drawbacks that could lead to misleading computed similarities. In particular, they consider only the overall tendencies, thereby failing to consider the contributions of each vector component. This is exemplified in Fig. 1 by the quantities in which two products, p_A and p_B, were purchased by two users. PCC cannot be calculated if a product is purchased in the same quantities by all users. Furthermore, it can take on negative values, which is inconsistent with our assumption of positive relevance scores for independent attributes. For the purpose of this paper, all negative values were considered to be zero. COS overestimates similarity by ignoring the total number of purchases. For example, cases (a) and (b) have the same COS similarity of 0.7; however, p_A and p_B were purchased by the same user only once in the former case, and 10 times in the latter case. Intuitively, we would expect the higher purchase frequency to indicate a higher similarity. Likewise, JAC considers only the number of users who purchased both products, irrespective of purchase quantities. For example, case (d) has a maximum value of 1 for the JAC measure, although both products were purchased in substantially different quantities.

We define a similarity measure that avoids the drawbacks of conventional similarities. The proposal is bound between 0 and 1, and assigns an increasing similarity as the total number of purchases increase. In particular, we focused more on products that had been purchased with the same frequency as that of the target product, and less on those for which the number of purchases disagree. We refer to this similarity as matched preference ratio (MPR).

$$
\begin{array}{lcl} & & \begin{array}{cc} u_1 & u_2 \end{array} \\ p_A & = & [0, \quad 1] \\ p_B & = & [1, \quad 1] \end{array}
$$

PCC is undefined

COS = 0.7

JAC = 0.5

(a)

$$
\begin{array}{lcl} & & \begin{array}{cc} u_1 & u_2 \end{array} \\ p_A & = & [0, \quad 10] \\ p_B & = & [10, \quad 10] \end{array}
$$

PCC is undefined

COS = 0.7

JAC = 0.5

(b)

$$
\begin{array}{lcl} & & \begin{array}{cc} u_1 & u_2 \end{array} \\ p_A & = & [0, \quad 1] \\ p_B & = & [1, \quad 0] \end{array}
$$

PCC = -1

COS = 0

JAC = 0

(c)

$$
\begin{array}{lcl} & & \begin{array}{cc} u_1 & u_2 \end{array} \\ p_A & = & [1, \quad 1] \\ p_B & = & [10, \quad 10] \end{array}
$$

PCC is undefined

COS = 1

JAC = 1

(d)

FIGURE 1. **Similarity scores between two product vectors, p_A and p_B. Each element of the vector corresponds to purchase quantities by two users, u_1 and u_2. PCC cannot be calculated in cases (a), (b), and (d), and takes on a negative value in case (c). COS overestimates similarities in cases (a), (b), and (d). JAC assigns the same similarity value in cases (a) and (b), and the maximum similarity value in case (d), because it considers exclusively whether a user bought both products or not.**

IV. PROPOSED METHOD

Considering a set of products $P = \{p_1, p_2, \cdots, p_N\}$ with known M-dimensional attribute vectors $a(p_n)$ (for each $n \in [1, N]$). The unknown attribute vector of a product q with known similarities to the products in P can be estimated as follows:

$$
a(q) = \frac{\sum_{i=1}^{N} w_i(q) a(p_i)}{\sum_{i=1}^{N} w_i(q)}, \tag{4}
$$

where w_i is the similarity of products p_i and q. In particular, we calculated this as follows:

$$
\begin{aligned}
w_i(q) &= \text{MPR}(p_i, q) \\
&= \frac{\sum\limits_{u \in U} I_a(p_i, q) c_{u,p_i} c_{u,q}}{\sum\limits_{u \in U} I_a(p_i, q) c_{u,p_i} c_{u,q} + \sum\limits_{u \in U} |c_{u,p_i} - c_{u,q}|}.
\end{aligned} \tag{5}
$$

Here, c_{u,p_i} represents the purchase count of product p_i by a user u in the set of all users U. $I_a(p_i, q)$ is an indicator function of agreement, defined as follows:

$$
I_a(p_i, q) = \begin{cases} 1 & \text{if } c_{u,p_i} = c_{u,q} \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}
$$

The numerator in (5) includes a nonzero term for every user who purchased product p_i in equal quantity as q. Therefore, the similarity score has a first-order relationship with the number of users who purchased both products in equal quantities. On the other hand, each term is the product of purchase amounts for p_i and q; this means that the relationship between the purchase quantity and the similarity score is quadratic. Concerning the denominator, the first sum ensures that the measure is normalized to values between 0 and 1. Meanwhile, the second sum penalizes the score in proportion to the number of users who purchased p_i and q in different quantities. The result is that two products are assigned a higher similarity score when more users purchase them in equal amounts and these quantities are large. In contrast, they are considered less similar if a large number of users purchased them in different quantities; particularly if the difference in purchase quantities is relatively large. These properties mean that the proposed similarity measure depends not only on the direction of the item vector, but also on the magnitude of its components (i.e., the purchase counts). This ensures that the proposal is free from the drawbacks of conventional similarity measures, outlined in Section III.

If products were classified into categories whose members can be assumed to share similar attributes, the model could be further improved by adding a term to (4); that takes into account the average of all products within the target product category. That is,

$$a(q) = \frac{\alpha \sum_{i=1}^{N} w_i(q) a(p_i)}{\sum_{i=1}^{N} w_i(q)} + \frac{(1-\alpha)}{|\mathrm{Cat}(q) \cap P|} \sum_{\rho \in \mathrm{Cat}(q) \cap P} a(\rho), \quad (7)$$

where α is a parameter estimated empirically, and acts as an adjusting factor that controls the relative contribution of each term; $\mathrm{Cat}(q)$ represents the set of products in the same category as q.

V. EXPERIMENTS

The data used in this research were provided by Rakuten Group's marketing research department, who conducted a survey among a sample of 18,000 Rakuten Ichiba users aged between 20 and 69 years old. The respondents were asked to evaluate 67 different brands of Japanese food products from different categories, such as chocolates, beers, and yogurts. Participation in the survey was voluntary and all respondents agreed to sign a consent form. Respondents were given a list of predefined attributes and asked to select those that characterized each product in the survey. If the respondent did not know the product or did not have any particular opinion about a product, they were encouraged to skip it. The attributes were carefully designed by marketing experts with the intent to describe the most representative features that reflect, at a fundamental level, the consumer sentiment towards a product. Marketing specialists narrowed down the attributes to a total of 39, selecting those that were considered to be the most insightful for the product research process. Examples of the attributes used are "It tastes sweet," "It feels healthy," and "It is refreshing." All products shared the same attributes. Table 2 gives an example of the survey format.

TABLE 2. Example of the survey format with products and attributes.

How do you feel about the following products?
(Check all that apply)

	Tastes sweet	Feels healthy	Is refreshing	...
Product A	□	□	□	□
Product B	□	□	□	□
Product C	□	□	□	□
...	□	□	□	□

The survey results were processed in two steps. In the first step, the replies were aggregated by product to obtain the total counts per attribute. We will refer to the counts as votes, as each count represents a respondent stating that an attribute is relevant to a product. Therefore, a count of zero implies that no respondent associated a particular product with the corresponding attribute. Because not every respondent would review every product, the total votes per product also varied depending on the popularity of the item being evaluated.

The second step involved the normalization of the votes by the total number of respondents per product to account for the effect of the popularity. This was necessary in order to be able to compare attributes between products in terms of the proportion of replies, instead of total number of votes. Without the normalization, the similarity score between products would be affected by the dominant influence of popular ones over those that are less known. That is, the normalization allows for a fair comparison of the attributes of different products, regardless of their degree of popularity.

This process allowed us to build a collection of product encodings, in which each product is associated with a vector of attributes. These vectors capture the strength with which each attribute was related to each product on a continuous scale from 0 to 1. For example, the "Is refreshing" attribute for a beer product with a value close to 1 means that most of the respondents found the attribute "Is refreshing" very relevant, whereas the same attribute for a chocolate product could have a value close to 0 (not very relevant), which means most of the respondents did not find the chocolate refreshing. Fig. 2 shows a visual representation of each product and its attributes. Table 3 presents a sample of the aggregated survey responses for three pairs of products from different categories and ten attributes. The attribute values of a given product are, in general, closer to those of other products in the same category, than to those in other categories. However, same-category products are not identical. That is, some level of distinctiveness exists within the categories; for example, Chocolate A is regarded to be notably more bitter than Chocolate B.

To build the product similarity matrix that would be used in later stages to estimate unknown attributes, we first created an item-user matrix similar to that in Table 4. We employed the purchase information provided by Rakuten Ichiba in the transaction history logs as an implicit measure of preference; a higher number of items bought indicates stronger preference. Therefore, each element in the item-user matrix is the

TABLE 3. Sample of aggregated survey responses for six different products and ten attributes [1].

Attributes \ Products	Beer A	Beer B	...	Chocolate A	Chocolate B	...	Yogurt A	Yogurt B
Tastes sweet	0.056	0.015	...	0.887	0.736	...	0.345	0.259
Tastes sour	0.012	0.001	...	0.002	0.016	...	0.158	0.133
Tastes spicy	0.03	0.148	...	0.008	0.002	...	0.016	0.012
Tastes bitter	0.198	0.407	...	0.238	0.016	...	0.02	0.022
Aroma is good	0.286	0.391	...	0.170	0.064	...	0.038	0.0436
...	...	...	...	...	...	...	...	...
Feels healthy	0.068	0.063	...	0.219	0.092	...	0.844	0.785
Feels delicate	0.067	0.149	...	0.113	0.256	...	0.126	0.107
Feels powerful	0.216	0.410	...	0.138	0.073	...	0.137	0.07
...	...	...	...	...	...	...	...	...
Is relaxing	0.744	0.523	...	0.478	0.43	...	0.370	0.292
Is refreshing	0.715	0.685	...	0.041	0.089	...	0.326	0.116

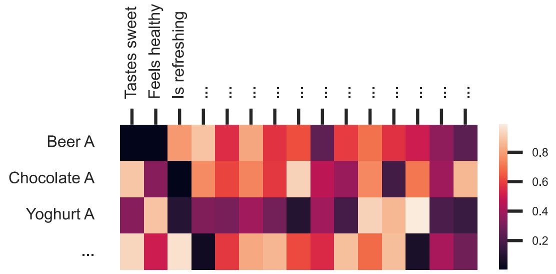

FIGURE 2. Visual representation of products and their attributes. Lighter colors indicate stronger relevance. According to the figure, the attribute "Is refreshing" is considered to be very relevant to "Beer A," whereas attributes "Is sweet" and "Is tasty" are not very relevant. Similarly, attributes "Is healthy" and "Is sweet" are considered to be relevant to "Yogurt A".

TABLE 4. *n* × *m* item-user matrix. Elements of the matrix represent the total number of items bought per user in a fixed period of time.

$$\begin{array}{c} \\ i_1 \\ i_2 \\ \vdots \\ i_n \end{array} \begin{array}{c} \begin{array}{cccc} u_1 & u_2 & \dots & u_m \end{array} \\ \begin{bmatrix} 0 & 2 & \dots & 1 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 8 & \dots & 0 \end{bmatrix} \end{array}$$

total number of items bought per user in a fixed period of time. Subsequently, the similarity between products could be calculated with pairs of product vectors from the item-user matrix and the equations presented in Section II-B and IV.

We evaluated the performance of different product attribute estimation models, each of them calculating attributes based on different similarity measures. Our experiments were performed several times using the leave-one-out cross-validation (LOOCV) technique [37]. The reported results are the average of all trials. We compared the result of our proposed similarity measure against three baselines: PCC, COS, and JAC.

Each model was built using a set of "base products," which served as a seed to estimate new attribute vectors. Therefore, in every trial, we split the set of survey products in two disjoint sets. The first set contained all products minus one; this was the "base products" set. The second set contained one single item, the "target product." This product was used to evaluate the accuracy of our estimations.

As we intended to measure how much our estimations deviated from the real values, we selected the mean absolute error (MAE) and root mean squared error (RMSE) as the evaluation metrics. In both cases, a lower value indicates a better performance; however, RMSE is more sensitive to deviations in individual components.

The procedure to calculate the evaluation metrics using LOOCV comprises the following steps:

For each product p_t ($t \in [1, N]$) in the survey:

1) Consider the set of all other products $P_{\text{base}} = P - \{p_t\}$ as known products, and p_t as the target product.
2) Compute an estimate for its attribute vector $\hat{a}(p_t)$ using (4).
3) Determine the error vector $\epsilon_t = |\hat{a}(p_t) - a(p_t)|$ as the vector of distances between the estimated and the actual attribute values.
4) Compute the MAE and RMSE per target product as follows:

$$\text{MAE}_t = \frac{\|\epsilon_t\|_1}{M} \tag{8}$$

and

$$\text{RMSE}_t = \frac{\|\epsilon_t\|_2}{\sqrt{M}} \tag{9}$$

5) Repeat steps 1-4 for all products in P.
6) Average the MAE and RMSE values obtained for all target items as follows:

$$\text{MAE} = \frac{1}{N}\sum_{t=1}^{N} \text{MAE}_t \tag{10}$$

and

$$\text{RMSE} = \frac{1}{N}\sum_{t=1}^{N} \text{RMSE}_t \tag{11}$$

Finally, our experiments considered the coefficient of determination r^2 as the square of the Pearson correlation

[1] The names of commercial brands for the products are not shown to avoid any potential trademark conflict. The attributes are English translations of the actual ones used in the Japanese-language survey.

coefficient. For two series x and y, this is given as:

$$r^2 = \left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \right)^2, \tag{12}$$

with x_i and y_i denoting the values in the series, while $\bar{x}$ and $\bar{y}$ stand for their respective average values. The sums run over all elements in the series.

TABLE 5. Performance metrics for various models using different similarity measures considering a purchase period of 50 days. The MPR model outperforms all other models, with the lowest MAE and RMSE.

Similarity	MAE	RMSE
MPR	0.040254	0.067911
JAC	0.0445531	0.0746888
COS	0.0449462	0.0766356
PCC	0.0596924	0.100329

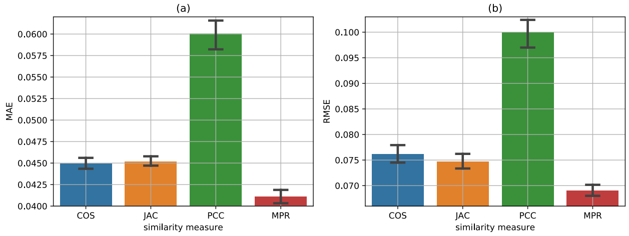

FIGURE 3. (a) MAE and (b) RMSE for various models using different similarity measures considering a purchase period of 50 days. The MPR model outperforms all other models, with the lowest MAE and RMSE.

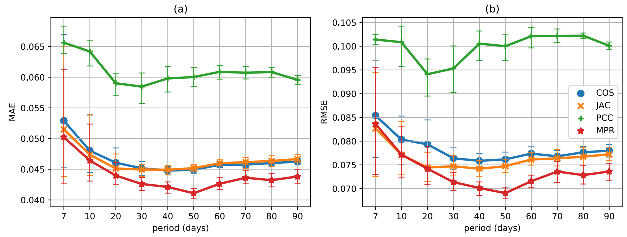

FIGURE 4. Performance comparison in terms of the (a) MAE and (b) RMSE of the baseline and proposed models using different purchase periods. The MPR model outperforms all baselines consistently.

VI. RESULTS

Table 5 shows MAE and RMSE values for each model using MPR and the three baseline measures. In this experiment, we considered the purchase history over a period of 50 days. This is also represented in Fig. 3 using bar plots with confidence intervals of 99%. The results show that MPR outperformed all other models. The plots in Fig. 4 show the performance metrics with 99% confidence intervals of the models built with purchase history periods of different lengths. The shortest period was one week, whereas the longest period was 90 days. It is possible to observe that the model built with our proposed similarity measure outperformed all baselines consistently. The low performance exhibited for shorter periods can be attributed to the lack of data to make reliable estimations. When the observations are limited to only a few days, not enough users make their purchases of the relevant products so as to reveal purchasing patterns. For the particular dataset used in our evaluation, the user diversity increased more than 10-fold by day 50 compared to the first 7 days for 11% of all the products. By day 50, the number of items bought is 6.49 times larger than the total items bought during the first week. For all models, the performance increased for larger periods, up to approximately 40 or 50 days; as the confidence intervals did not overlap, the difference was statistically significant. At that point, the item-user matrix provided a reliable estimate of similarity.

Fig. 5 shows the performance of various models using different purchase periods. Each bar represents the proportion of times each model achieved the lowest MAE per trial. It is possible to see that, for all periods, the model relying on MPR similarity achieved the lowest error with the highest frequency.

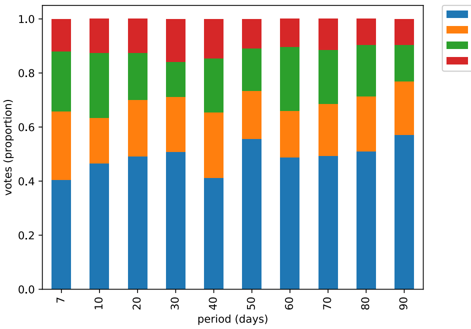

FIGURE 5. Performance of different models regarding the proportion of times each model achieved the lowest MAE per trial. For all trials, the MPR model achieved the lowest error with the highest frequency.

The experiments presented so far use the attribute information of similar products to compute estimations, regardless of whether they belong to the same category or not. Products from different categories can contribute to the predicted attribute values as long as they are considered similar. To complement this analysis, we investigated the impact of using the statistics of the target product's category on the performance of the estimation. The products were manually classified into categories by product marketing specialists. However, this is typically costly and unfeasible for large catalogues, for which classification errors usually can be be expected. To simulate this, we performed a random shuffling of two products per category. Fig. 6 shows the MAE (a) and RMSE (b) of two models, with a one-month purchase period, averaged over 10 trials. The model that uses MPR similarity is plotted as a reference and it remains constant because it does not depend on α. In contrast, the performance of the hybrid model (described earlier in (7)) improves up to some point as $1 - \alpha$ increases, and thereafter deteriorates and becomes worse than the reference performance. This indicates that

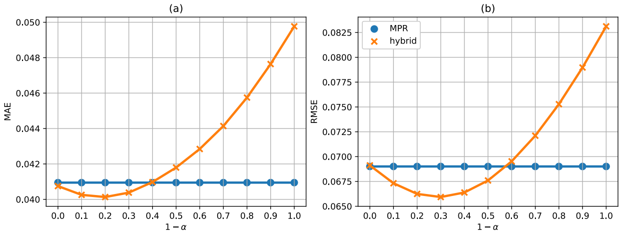

FIGURE 6. **Effect of category information on the performance of the estimation. The performance of the hybrid model ((7)) improves and worsens as a function of α. When category information is available, some improvement in performance can be achieved by combining both similarity and category information.**

TABLE 6. **Numerical comparison between the actual and estimated attribute values for "Beer A".**

Attributes	Actual	Estimated	Absolute error
Tastes sweet	0.056	0.033	0.022
Tastes sour	0.012	0.010	0.001
Tastes spicy	0.03	0.077	0.047
Tastes bitter	0.198	0.244	0.046
Aroma is good	0.286	0.197	0.088
...	...	...	...
Feels healthy	0.068	0.062	0.006
Feels delicate	0.067	0.075	0.008
Feels powerful	0.216	0.140	0.075
...	...	...	...
Is relaxing	0.744	0.616	0.127
Is refreshing	0.715	0.694	0.020

when category information is available, some improvement in performance can be achieved by combining both similarity and category information. However, the category information alone is not sufficient to produce accurate estimations.

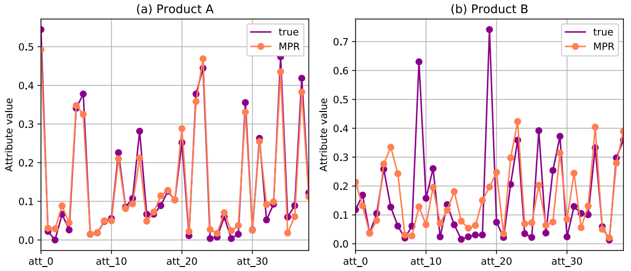

FIGURE 7. **Estimated attribute values of two products using the MPR model versus the true attribute values for two products. (a) Example of an estimation with low error. (b) Example of an estimation with high error.**

Fig. 7 demonstrates the estimated attribute values of two products using the MPR model versus the true attribute values. Product A and product B are examples of estimations with low and high error, respectively. It is apparent that the estimations of product A follow the true values very closely. However, this is not the case for Product B. Product B is an example of a product with unique attributes, even for its own category. Our model failed to give close estimations for such cases because the MPR model computed the unknown attributes as a linear combination of those of similar items. When there were no examples of similar items, the estimations were unsatisfactory. Additionally, Table 6 shows a numerical comparison between the actual and estimated attribute values of "Beer A" previously presented in Table 3.

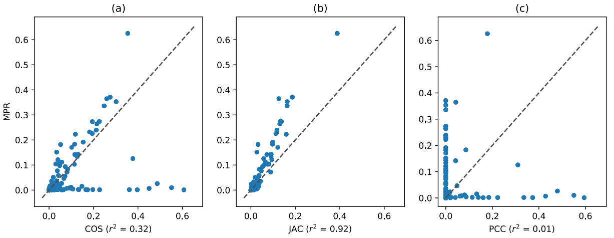

FIGURE 8. **Correlation between MPR similarity measure and the baseline similarity measures. (a) MPR assigns low similarity in a subset of products, whereas COS varies widely. (b) MPR and JAC are considerably correlated, however, most points lie above the diagonal in the plot. This means that JAC tends to assign lower similarity values when compared to MPR (i.e. it underestimates similarity values compared to MPR). (c) MPR and PCC are uncorrelated and appear to be looking at different aspects of the data.**

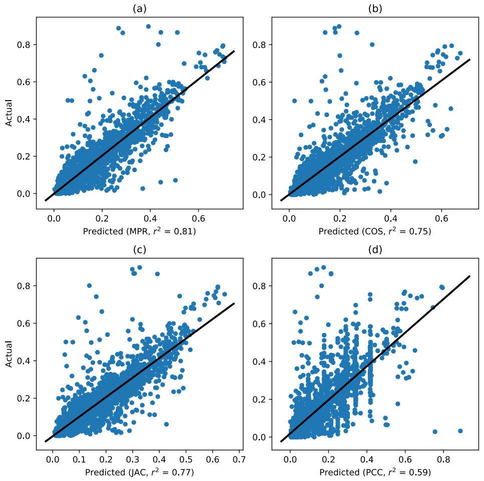

FIGURE 9. **Predicted versus actual scatter plots for each similarity measure. Each point is the value of an attribute for every product. MPR achieves the highest r^2 among all models.**

Fig. 8 shows the correlations between our proposed similarity measure and the baseline similarity measures, with the coefficient of determination r^2 calculated according to (12). It is evident that JAC is correlated; however, it tends to assign lower similarity values than MPR. The correlation between JAC and MPR arises from the fact that most items were purchased in smaller quantities. One of the main features of MPR is the weight it gives to instances where items were purchased with higher frequency, or when the difference between purchase amounts was larger. This gives MPR an edge over JAC. COS exhibits a similar trend, but there is a subset of products for which MPR gives a low similarity, whereas the COS measure varies widely. This corresponds to products that were purchased in smaller quantities; therefore, MPR assigns a lower weight to them, whereas COS considers only the angle between row-vectors of the item-user matrix, which can take any value, irrespective of purchase numbers.

Finally, PCC is uncorrelated; it appears that both measures focus on independent features of the data.

Fig. 9 shows a scatter plot of the predicted values against the actual values for each attribute and every product. The estimated values of a perfect model would all fall on the diagonal. The coefficient of determination r^2 is calculated as shown in (12). The MPR model achieves the highest r^2 among all models, meaning that the error in its estimations is lower than that of the other models.

VII. CONCLUSION

In this paper, we proposed a method for estimating product attribute vectors based on survey data and purchase history. This is particularly useful for practical applications in marketing or e-commerce where resources are limited and acquiring additional data to train conventional machine learning models is time-consuming and expensive. The proposed method involves estimating unknown attribute vectors as the weighted average of those of known products. The values of those weights depend on the degree of similarity between the target product and the base products, i.e., products with known attribute values. Furthermore, we proposed a new similarity measure that is designed based on domain-specific assumptions. Our approach relies on the total number of items bought as an implicit measure of preference. The proposed method assigns a larger similarity score to products that were bought in equal amounts, and a smaller one as the difference in purchase amounts increases. This feature allows us to overcome the weaknesses of the conventional similarity measures. The proposed method was applied to data surveyed by Rakuten Group's marketing research department and the purchase history from Rakuten Ichiba. The results obtained indicate that the proposed method outperforms the conventional approaches in terms of MAE and RMSE.

However, our study was limited to the kind of products frequently purchased on Rakuten Ichiba. These consist mainly of daily use products, such as food and beverages. For this reason, it is unknown whether the proposal can remain effective for product catalogues of a different kind, such as cars or electronic equipment which are not purchased frequently and in high numbers. Further, the dataset does not fully sample all types of customers. It is limited to Japan, and biased towards the age and social groups that tend to make online purchases. Nevertheless, the results show the practical applicability of the proposal in the case study. Further investigation of its generality is considered to be a promising line for future research. Other potential extensions to this study include the application of the proposed similarity measure in collaborative filtering (e.g., to predict top-N recommendations) and its extension to account for products with unique features.

ACKNOWLEDGMENT

The authors would like to thank J. Trevino for insightful discussions, and Y. Kuno along with the Rakuten Insight team for providing the survey data.

REFERENCES

[1] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "Opentag: Open attribute value extraction from product profiles," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, 2018, pp. 1049–1058.

[2] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano, "Text mining for product attribute extraction," *ACM SIGKDD Explor. Newslett.*, vol. 8, no. 1, pp. 41–48, Jun. 2006.

[3] S. Raju, P. Pingali, and V. Varma, "An unsupervised approach to product attribute extraction," in *Proc. 31th Eur. Conf. IR Res. Adv. Inf. Retr.* Berlin, Germany: Springer-Verlag, Apr. 2009, pp. 796–800.

[4] A. More, "Attribute extraction from product titles in ecommerce," in *Proc. Enterprise Intell. Workshop*, Aug. 2016, pp. 1–5.

[5] S. Vajjala, B. Majumder, H. Surana, and A. Gupta, *Practical Natural Language Processing: A Pragmatic Approach to Processing and Analyzing Language Data*. Newton, MA, USAO'Reilly Media, 2020.

[6] Z. Kozareva, Q. Li, K. Zhai, and W. Guo, "Recognizing salient entities in shopping queries," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 107–111.

[7] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, "Neural architectures for named entity recognition," in *Proc. Conf. North Amer. Chapter Assoc. for Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 260–270.

[8] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany, Aug. 2016, pp. 1064–1074.

[9] E. Rosch, "Principles of categorization," in *Cognition Categorization*. Hillsdale, NJ, USA: Erlbaum, 1978, pp. 27–48.

[10] C. H. Brooks and N. Montanez, "Improved annotation of the blogosphere via autotagging and hierarchical clustering," in *Proc. 15th Int. Conf. World Wide Web*, New York, NY, USA, 2006, pp. 625–632.

[11] S. Golder and B. Huberman, "The structure of collaborative tagging systems," *J. Inf. Sci.*, vol. 32, p. 15, Sep. 2005.

[12] M. Guy and E. Tonkin, "Folksonomies: Tidying up tags?" *D-Lib Mag.*, vol. 12, no. 1, Jan. 2006.

[13] S. Hayman, "Folksonomies and tagging: New developments in social bookmarking," in *Proc. Ark Group Conf., Developing Improving Classification Schemes*. Sydney, NSW, Australia, Jun. 2007, pp. 1–5.

[14] M. Kipp and D. G. Campbell, "Patterns and inconsistencies in collaborative tagging systems : An examination of tagging practices," *Amer. Soc. Inf. Sci. Technol.*, vol. 43, pp. 1–8, Oct. 2007.

[15] K. Probst, R. Ghani, M. Krema, A. Fano, and Y. Liu, "Extracting and using attribute-value pairs from product descriptions on the web," in *Web to Social Web: Discovering Deploying User Content Profiles*. Berlin, Germany: Springer Berlin, 2007, pp. 41–60.

[16] L. Bing, T.-L. Wong, and W. Lam, "Unsupervised extraction of popular product attributes from E-Commerce Web sites by considering customer reviews," *ACM Trans. Internet Technol.*, vol. 16, no. 2, pp. 1–17, Apr. 2016.

[17] S. Kovelamudi, S. Ramalingam, A. Sood, and V. Varma, "Domain independent model for product attribute extraction from user reviews using wikipedia," in *Proc. 5th Int. Joint Conf. Natural Lang. Process.*, Nov. 2011, pp. 1408–1412.

[18] F. Tang, L. Fu, B. Yao, and W. Xu, "Aspect based fine-grained sentiment analysis for online reviews," *Inf. Sci.*, vol. 488, pp. 190–204, Jul. 2019.

[19] L. Zhang and B. Liu, *Sentiment Analysis Opinion Mining*. Boston, MA, USA: Springer, 2017, pp. 1152–1161.

[20] J. Vig, S. Sen, and J. Riedl, "The tag genome: Encoding community knowledge to support novel interaction," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 1–44, Sep. 2012.

[21] D. Eck, P. Lamere, T. Bertin-Mahieux, and S. Green, "Automatic generation of social tags for music recommendation," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, 2007, pp. 385–392.

[22] A. Ulges, C. Schulze, D. Keysers, and T. M. Breuel, "A system that learns to tag videos by watching youtube," in *Computer Vision Systerm*. Berlin, Germany: Springer, 2008, pp. 415–424.

[23] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social tag prediction," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, New York, NY, USA, 2008, pp. 531–538.

[24] M. M. Deza and E. Deza, *Encyclopedia of Distances*, 3rd ed. Berlin, Germany: Springer-Verlag, 2014.

[25] G. Guo, J. Zhang, and N. Yorke Smith, "A novel Bayesian similarity measure for recommender systems," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 2619–2625.

[26] C. C. Aggarwal, *Recommender Systems: The Textbook*, 1st ed. Cham, Switzerland: Springer, 2016.

[27] S. S. Choi, S. H. Cha, and C. C. Tappert, "A survey of binary similarity and distance measures," *J. Systemics, Cybern. Informat.*, vol. 8, pp. 43–48, Nov. 2009.

[28] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Models Methods Appl. Sci.*, vol. 1, no. 4, pp. 300–307, 2007.

[29] A. Agarwal, M. Chauhan, and Ghaziabad, "Similarity measures used in recommender systems : A study," *Int. J. Eng. Technol. Sci. Res.*, vol. 4, pp. 2394–3386, Jun. 2017.

[30] H. J. Ahn, "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem," *Inf. Sci.*, vol. 178, no. 1, pp. 37–51, Jan. 2008.

[31] Suryakant and T. Mahara, "A new similarity measure based on mean measure of divergence for collaborative filtering in sparse environment," *Procedia Comput. Sci.*, vol. 89, pp. 450–456, 2016.

[32] L. A. Hassanieh, C. A. Jaoudeh, J. B. Abdo, and J. Demerjian, "Similarity measures for collaborative filtering recommender systems," in *Proc. IEEE Middle East North Afr. Commun. Conf. (MENACOMM)*, Apr. 2018, pp. 1–5.

[33] Z. Tan and L. He, "An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle," *IEEE Access*, vol. 5, p. 27211–27228, 2017.

[34] P. H. Guzzi, M. Mina, C. Guerra, and M. Cannataro, "Semantic similarity analysis of protein data: Assessment with biological features and issues," *Briefings Bioinf.*, vol. 13, no. 5, pp. 569–585, Dec. 2011.

[35] S. Jain and G. D. Bader, "An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology," *BMC Bioinf.*, vol. 11, no. 1, p. 562, Nov. 2010.

[36] S. Wan, M.-W. Mak, and S.-Y. Kung, "HybridGO-loc: Mining hybrid features on gene ontology for predicting subcellular localization of multi-location proteins," *PLoS ONE*, vol. 9, no. 3, Mar. 2014, Art. no. e89545.

[37] C. Sammut and G. I. Webb, *Encyclopedia Machine Learning*. Boston, MA, USA: Springer, 2010, pp. 600–601.

PATRICIA ORTAL (Member, IEEE) is currently pursuing the Ph.D. degree in information science with Nagoya University.

She has been working as a Research Scientist with the Rakuten Institute of Technology, Tokyo, Japan, since 2016. Her research interests include artificial intelligence, recommender systems, and behavior analysis.

MASATO EDAHIRO (Member, IEEE) received the Ph.D. degree in computer science from Princeton University, Princeton, NJ, USA, in 1999.

He joined NEC Corporation in 1985, where he worked in the research center for 26 years. He is currently a Professor with the Graduate School of Information Science, Nagoya University, Nagoya, Japan. His research interests include graph and network algorithms and software for multi-core and many-core processors.