Inferring Win–Lose Product Network from User Behavior

Shuhei Iitsuka
The University of Tokyo
Hongo 7-3-1
Bunkyo, Tokyo, Japan
iitsuka@weblab.t.u-tokyo.ac.jp

Kazuya Kawakami
University of Oxford
Wolfson Building, Parks Road
Oxford, UK
kazuya.kawakami@cs.ox.ac.edu

Seigen Hagiwara
Recruit Marketing Partners Co., Ltd.
Kyobashi 2-1-3
Chuo, Tokyo, Japan
seigen@r.recruit.co.jp

Takayoshi Kawakami
Industrial Growth Platform, Inc.
Marunouchi 1-9-2
Chiyoda, Tokyo, Japan
t.kawakami@igpi.co.jp

Takayuki Hamada
IGPI Business Analytics & Intelligence, Inc.
Marunouchi 1-9-2
Chiyoda, Tokyo, Japan
t.hamada@igpi.co.jp

Yutaka Matsuo
The University of Tokyo
Hongo 7-3-1
Bunkyo, Tokyo, Japan
matsuo@weblab.t.u-tokyo.ac.jp

ABSTRACT

Various data mining techniques to extract product relations have been examined, especially in the context of building intelligent recommender systems. Most such techniques, however, specifically examine co-occurrences of browsed or purchased products on e-commerce websites, which provide little or no useful information related to the direct relation of superiority or the factor which forms that superiority. For marketers and product managers, understanding the competitive advantages of a given product is important to consolidate their product differentiation strategies.

As described in this paper, we propose a *win–lose relation*, a new product relation analysis method that retrieves the superiority relation between competitive products in terms of product attractiveness. Our proposed method uses the difference between user browsing and purchasing behaviors, assuming that a purchased product is superior to products that are browsed but not purchased. We also propose superiority factor analysis to examine keywords that represent the superiority factor by mining product reviews. We evaluate our methods using an actual dataset from Zexy, the largest wedding portal website in Japan. Our experimental evaluation revealed that our proposed method can estimate actual user preferences observed from a user study using only log data. Results also show that our proposed method raises the accuracy of superiority factor extraction by around 17% by considering the win–lose relation of products.

CCS CONCEPTS

•Information systems →Web log analysis; *Electronic commerce;* Data mining;

KEYWORDS

product network, e-commerce, review mining

WI '17, Leipzig, Germany

DOI: 10.1145/3106426.3106502

ACM Reference format:
Shuhei Iitsuka, Kazuya Kawakami, Seigen Hagiwara, Takayoshi Kawakami, Takayuki Hamada, and Yutaka Matsuo. 2017. Inferring Win–Lose Product Network from User Behavior. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 8 pages.
DOI: 10.1145/3106426.3106502

1 INTRODUCTION

The e-commerce market is expanding along with the increasingly wider use of the internet. As the market expands widely, various data mining techniques are becoming used to extract useful information for e-commerce marketing. Starting from market basket analysis, extracting association rules of products has been studied extensively [9]. By combination with a network analysis approach, they enabled product marketers to extract interesting information in the form of product networks [2]. Most such methods, however, provide no information about their mutual superiority relations in terms of product attractiveness and the reason why the superiority is formed. Understanding the competitive advantages and relative benefits of a product is important for marketers to consolidate their branding strategies and to differentiate their products from those of their competitors. Nevertheless, few data mining techniques have been proposed to address this important problem.

We propose a new product relation analysis method that specifically examines the superiority relation among substitutable products, which we call the *win–lose relation.* Our proposed method uses the difference between users' browsing and purchasing behaviors on e-commerce websites. The methodology is simple: we define a win–lose relation from products that are browsed but not purchased to ones purchased by a given user. By aggregating the win–lose relations, one can define the *win–lose network* of products, which provides an overview of their superiority relations. We also propose *superiority factor analysis*, which examines keywords that represent the superiority factor by mining product reviews.

We present an overview of the proposed method in Figure 1 with the example of substitutable camera products. Each product has different selling points and competitors. The *stylish* design might be the selling point for a compact camera against its competitive compact cameras, although its *size* might be the point of appeal

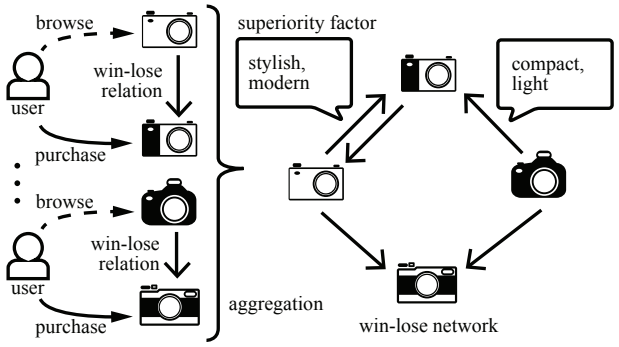


Figure 1: Overview of win–lose relation and superiority factor.

against professional cameras. Our proposed method enables visualization of the dynamics of competition and keywords that represent selling points against a specific competitor.

In analysis results, we used an actual dataset from the largest Japanese wedding portal website, Zexy[1], which is associated with thousands of wedding venues in Japan. Assuming the wedding venues as products, we analyzed the competitive and win–lose relations and visualized them in the form of a network. Using reviews of the wedding venues, we also conducted superiority factor analysis and examined the superiority factor.

For experimental evaluation, we evaluated that our proposed method can make a good estimation of the actual user behavior and preference extracted from Zexy's user survey results. Results showed a significant correlation between them, which means that our proposed method can estimate the actual win–lose relation from log data as long as we accept the definition of a win–lose relation we propose in this paper. We also demonstrated that our proposed method can infer superiority factor keywords by improving accuracy by around 17% compared to the baseline method, which does not consider win–lose relation.

The contributions of this paper are summarized as follows.

- We proposed a new data mining method to analyze a superiority relation in terms of product attractiveness. E-commerce website owners can estimate users' preferences from log data, which enables them to plan effective marketing and promotion strategies.
- We proposed a text mining method to analyze the superiority factor using information from product reviews. Product managers can use this for their differentiation strategies. Researchers can introduce widely diverse natural language processing methods for additional and sophisticated investigations.
- We evaluated log data as a substitute for user survey results in terms of estimating the actual superior relations of products and the superiority factor. Product managers and e-commerce website owners can benefit by saving the costs of conducting user surveys, which can be huge costs involving outlays for questionnaire distribution and data collection.

[1]Zexy http://zexy.net/

The remainder of this paper is organized as explained below. We introduce related works in Section 2 and our proposed method in Section 3. Section 4 introduces some application examples with an actual dataset. Section 5 describes the evaluation experiment. After the discussion presented in Section 6, we conclude this paper in Section 7.

2 RELATED WORKS

Product relations have been studied extensively by both microeconomics and data mining research communities. In the field of microeconomics, understanding of product relations is regarded as necessary for product managers to make marketing mix decisions [6]. In consumer theory, product relations are categorized into two kinds: *substitute* or *complementary*. A mobile device, for example, manufactured by another brand is a substitute product for a mobile device, although a mobile charger is a complementary product. Product substitutability and complementarity have long been common means of perceiving product relations [10]. This idea has been applied to widely diverse markets to ascertain consumer behavior [12].

Product relation analysis has been imported to e-commerce marketing, especially for implementation of advanced recommender systems. Zheng implemented a recommender system to meet user needs in different purchase stages considering product substitutability and complementarity [13]. McAuley defined co-purchased products as complementary products and co-browsed products as substitute products, and proposed a method to infer a network of complementary and substitute products, which is useful to generate context-relevant recommendations [7]. Product relation analysis have been examined to improve e-commerce marketing as described above, but few studies have specifically examined directional and adversarial relation between substitute products, which can be useful to plan product branding and differentiation strategy among competitors. Win–lose relations are a new form of product network that examines the superiority relation between substitute products by using both browsing and purchasing behaviors on e-commerce websites.

Several studies have proposed the use of product networks for e-commerce marketing. Hao et al. defined product networks based on their association rule and visualized them on a spherical surface [2]. Various characteristics of the product network have been studied by application of network analysis methods and by assuming it as a social network [8, 9]. Zinoviev proposed a method to discover users' contexts underlying purchasing behavior from product networks [14]. Product networks have been studied from various aspects as described above, but few researchers have examined the directed network of a superiority relation in product attractiveness such as a win–lose relation.

Review analysis is a major research field related to e-commerce marketing [11]. Especially, review summarization and review selection have been studied extensively to extract useful information from thousands of user reviews to help users make good decisions [3, 5]. Archak et al. proposed a method to extract important product features from product reviews by particularly addressing

the included noun phrases and associated adjectives [1]. That proposal inspired us to formalize the superiority factor analysis, which uses noun phrases included in product reviews.

3 PROPOSED METHOD

3.1 Formalization of Product Relations

When users visit an e-commerce website and consider purchasing products, they might browse multiple products to make comparisons. This tendency might become most apparent when they are making large purchases. We can define a subset of products browsed by the given user. Also, we can assume that the member products are in competition because they are in consideration for purchase by the same user who has specific needs. Therefore, we assume that these products share a *competitive relation* when they are browsed simultaneously by the same user.

As a result of the transaction, products are separable into three categories: products not browsed, products browsed but not purchased, and products purchased. The products browsed but not purchased are explicitly declined by the user as a result of consideration, which means that they are inferior to the purchased products in terms of product attractiveness. Therefore, we can define a superiority relation of product attractiveness among these products, naming the products browsed but not purchased as *loser products* and designating the purchased products as *winner products*. We designate this directed relation as a *win–lose relation*, connecting loser products to winner products. We define no win–lose relation among loser products or winner products.

Consider a set of products P handled on an e-commerce website and a user $u \in U$ who browses a set of products $P_u^{browse} \subset P$ and purchases a set of products $P_u^{purchase} \subset P$. In this case, we can define winner products as a set of purchased products $P_u^{win} = P_u^{purchase}$ and loser products as a subtraction of purchased products from browsed products $P_u^{lose} = P_u^{browse} \setminus P_u^{purchase}$. The competitive relation is defined among browsed products, whereas the win–lose relation is defined from any loser product to any winner product. Therefore, the set of competitive relations for user u is denoted as $R_u^C = \{\{p_i, p_j\} | \forall p_i, p_j \in P_u^{browse}, i \neq j\}$. The set of win–lose relations is $R_u^{WL} = \{(p_i, p_j) | \forall p_i \in P_u^{win}, \forall p_j \in P_u^{lose}\}$.

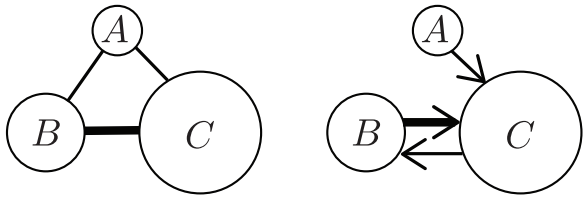
Assuming that an e-commerce website is handling a set of products $P = \{p_1, \cdots, p_5\}$, for example, and assuming that a user purchased a subset of products $P_u^{purchase} = \{p_1, p_2\}$ after browsing a subset of products $P_u^{browse} = \{p_1, p_2, p_3, p_4\}$, then the winner product set is $P_u^{win} = \{p_1, p_2\}$. The loser product set is $P_u^{lose} = \{p_3, p_4\}$, which presents six competitive relations of $R_u^C = \{\{p_1, p_2\}, \{p_1, p_3\}, \{p_1, p_4\}, \{p_2, p_3\}, \{p_2, p_4\}, \{p_3, p_4\}\}$ and four win–lose relations $R_u^{WL} = \{(p_1, p_3), (p_1, p_4), (p_2, p_3), (p_2, p_4)\}$.

3.2 Product Network

Visualizing the product relation as a network is useful to survey the characteristics of their competitive relations. By applying network analysis methods to this problem, one might identify clusters of competition or an interesting set of win–lose circulating relations similar to rock–paper–scissors. In this subsection, we introduce the *competitive network*, which represents competitive relations between products and the *win–lose network*, which represents superiority relations in product attractiveness.

Table 1: Example of product relations.

User u	Loser P_u^{lose}	Winner P_u^{win}
u_1	p_B	p_C
u_2	p_A, p_B	p_C
u_3	p_C	p_B



(a) Competitive network. **(b) Win–lose network.**

Figure 2: Example of a product network.

The competitive network is an undirected graph $G^C = (P, R^C)$, which consists of the nodes of products $p \in P$ and the edges of competitive relation $R^C = \bigcup_{u \in U} R_u^C$. We define that the weight of the edge between node p_i and node p_j is represented as the number of times that the competitive relation is defined between them, i.e., $w_{ij}^C = |\{u \in U | \{p_i, p_j\} \in R_u^C\}|$. However, the win–lose network is a directed graph $G^{WL} = (P, R^{WL})$, which consists of the nodes of products $p \in P$ and the edges of win–lose relation $R^{WL} = \bigcup_{u \in U} R_u^{WL}$. Similarly, the weight of the edge from node p_i to node p_j is the number of times that the win–lose relation is defined in that direction, i.e., $w_{ij}^{WL} = |\{u \in U | (p_i, p_j) \in R_u^{WL}\}|$.

For example, we assume three users who perceive their own loser products and winner products as described in Table 1. In this case, the competitive network and the win–lose network are expressed as shown in Figure 2. The edge thickness reflects its weight. The node size represents the number of users who purchased the product.

3.3 Superiority Factor Analysis

Interpreting the product relation is useful for e-commerce marketing activities, but marketers might also want to clarify the decisive factors which differentiate their own products from those of specific competitors. E-commerce website owners can use this information for recommendation by considering a given user's preference. Product managers might conceive a differentiation plan for their next product design. Understanding the superiority factors might enable them to ascertain which strategy is better: reinforcing their strength or repairing their weaknesses. We introduce the *superiority factor analysis* method, which extracts keywords, or *factor words*, that represent superiority factors to specific competitors using product reviews.

A simple means of extracting the factor word is calculating the term frequency of noun phrases that represent the product characteristics in product reviews. We can also use some techniques to filter out common terms or stop words. These techniques might

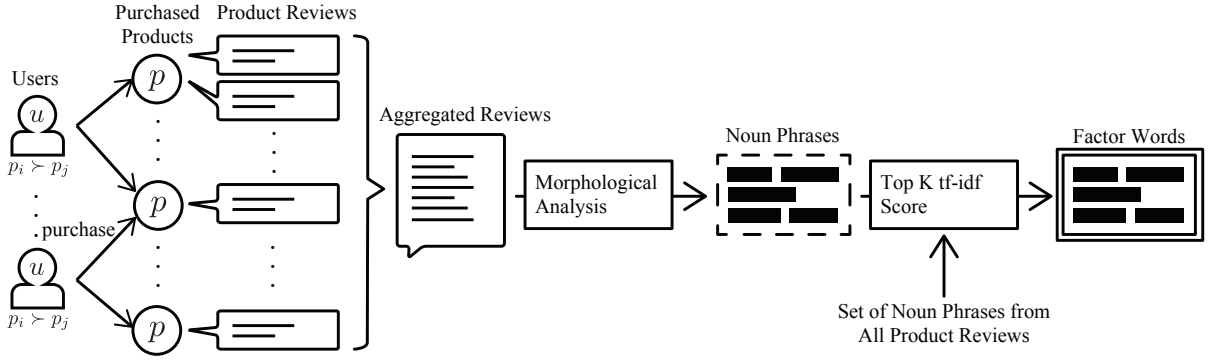


Figure 3: Overview of superiority factor analysis method.

be helpful to elucidate the product's general selling points in the market, but our proposed method differs from them in the point of retrieving the superiority factors against a specific competitor using a directed win–lose relation among them. We assume that each user has specific needs and preferences that are satisfied by the purchased products. Standing on that assumption, the reviews of products purchased by a set of users who support a specific win–lose relation might represent the point of view which constructed the win–lose relation. Therefore, we extract the factor words of a specific win–lose relation by application of a keyword extraction method to the reviews of products purchased by a user group that supports the win–lose relation.

We present an overview of the superiority factor analysis in Figure 3 with an example of extracting the factor words between winner product p_i and loser product p_j. We first define the set of users for whom winner products include product p_i and loser products include product p_j. We denote such users, who support the win–lose relation $p_i \succ p_j$, as $U_{p_i \succ p_j} = \{u \in U | p_i \in P_u^{win}, p_j \in P_u^{lose}\}$. Subsequently, we aggregate the product reviews of their purchased products, denoted as $P_{p_i \succ p_j} = \{p \in P_u^{purchase} | u \in U_{p_i \succ p_j}\}$, to produce a single aggregated review. We apply morphological analysis to the aggregated review and process it into a set of noun phrases.

Finally, we calculate the tf–idf score of every noun phrase in the processed review text by taking all product reviews as the corpus. We extract top K noun phrases with the highest tf–idf score as the factor words of the given win–lose relation. The tf–idf score of word w is calculated as a product of the term frequency tf_w and the inverse document frequency idf_w. The term frequency tf_w denotes the frequency at which the given word w appears in the review text in interest. The inverse document frequency idf_w is calculated as $idf_w = \log(N/N_w)$, where N denotes the total number of product reviews and N_w denotes the number of product reviews that include the word w.

4 ANALYSIS RESULTS

We apply our proposed method to the largest Japanese wedding portal website, Zexy, and present results to demonstrate the use of our proposed method with the actual dataset. Our proposed method works better when users browse many products for comparison and make careful considerations for purchase because our proposed method relies on differences between users' browsing and purchasing behaviors. The more products are browsed, the more product relations are defined, which makes the analysis more sophisticated. Wedding venues might fit to our analysis method well because reserving a venue is an important contract that demands careful consideration for most people. Therefore, we selected this website as the subject of this analysis. Zexy is not an e-commerce website in the strict sense, but we assume that it is in a broader sense because users take conversion actions such as making reservations for wedding venue tours and sending inquiries as a result of comparison on this website. We apply our proposed method by assuming browsing activity on wedding venues as browsing behavior and by assuming making reservations for venue tours as purchasing behavior.

The dataset comprises access log data collected during January 1, 2012 – October 31, 2012. Each record consists of an anonymized session ID, the browsed page URL which can be decoded as ID of the wedding venue, and the flag which indicates that the reservation for a venue tour is completed on that venue. Hereinafter, we assume this anonymized session ID as the identifier of users. During this time period, we found that around several million sessions are browsing the wedding venues, and that several tens of thousand sessions are making reservations. For analysis, we use only those sessions which include browsing and reservation actions.

Figure 4 presents the competitive network of wedding venues throughout Japan. Each node represents a wedding venue. Each edge represents their mutual competitive relation. The node size denotes the number of times the given product is purchased. The edge thickness represents the weight of the edge. The node color shows the region of Japan in which the wedding venue is located. For simplicity, we removed nodes from the illustration if its number of browsed times is less than a threshold. As a result, the figure shows approximately 1,000 nodes and 75,000 edges. The network is laid out based on a kinetic model, which locates connected nodes closer, simulating the spring force between them and unconnected nodes simulating the magnetic force [4]. As this figure shows, they form some clusters of competition. The color segments match well with the cluster segments. We ascertain that the competition takes place primarily in each region.

Figure 5 portrays the win–lose network of 10 selected wedding venues $(p_A, \cdots, p_J)$ in Tokyo. We selected these popular wedding

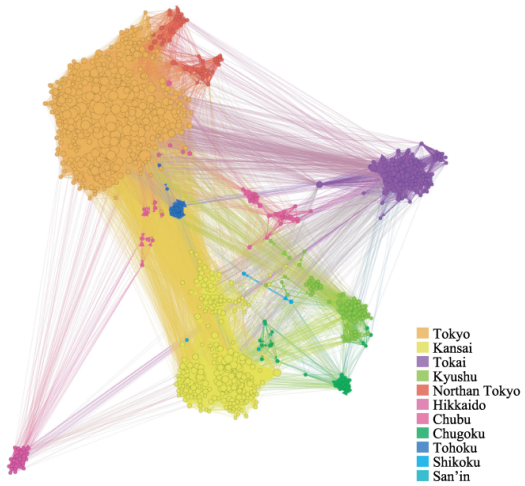


Figure 4: Competitive network of wedding venues throughout Japan.

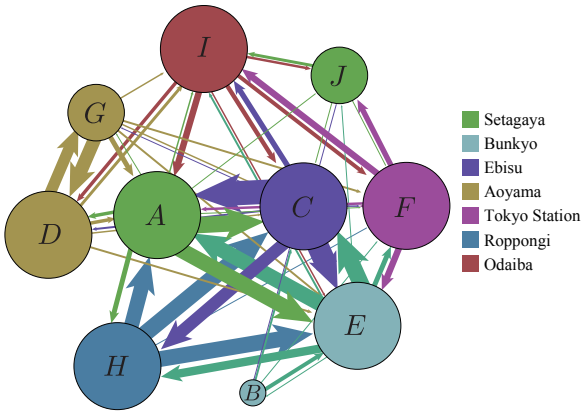


Figure 5: Win–lose network of selected wedding venues in Tokyo (colored by area).

venues because they are more likely to have sufficient data for analysis than other venues. Results show that edges among wedding venues p_A, p_C, p_E, and p_H are especially bold, which means that they share a strong competitive relation. When we specifically examine the win–lose relation from wedding venue p_H to wedding venue p_A, the outbound edge is much thicker than the inbound edge, which means that wedding venue p_H is inferior to wedding venue p_A in terms of product attractiveness. Therefore, the e-commerce website owner can expect that users tend to make conversion actions on wedding venue p_A rather than wedding venue p_H when they are browsed during the same session.

Finally, we conducted superiority factor analysis of wedding venues p_A, p_H, and p_J, which have large node sizes and a competitive relation among the 10 selected wedding venues. We use several thousand product reviews for the wedding venues in Tokyo as the corpus collected on the website. We treat every product review as a document and calculated tf–idf values of included noun phrases and

Table 2: Factor words of wedding venue p_H against wedding venue p_A and p_J.

Against p_A	Against p_J
ceremony	Japanese dish (kaiseki)
garden	Japanese-style room
banquet	ceremony
solemnity	garden
photograph	bus

extracted factor words. We conducted morphological analysis of review texts in Japanese using Mecab[2], which separates a sentence into a list of words with part-of-speech tagging information.

Table 2 shows top five factor words of wedding venue p_H against wedding venues p_A and p_J. Results show that product p_H has different superiority factors against each competitor. Against wedding venue p_A, *garden* and *banquet* arise as decisive factors. Both wedding venues p_A and p_H offer Japanese traditional style weddings, but wedding venue p_H apparently has superiority in terms of this characteristic. In addition, the photograph service quality is one important feature that users who prefer wedding venue p_H cannot compromise. Against wedding venue p_J, *Japanese dish (kaiseki)* and *Japanese-style room* come to the top of the list. Because wedding venue p_J is a hotel, the Japanese-style ceremony itself seems to be the strength of wedding venue p_H against wedding venue p_J. The transportation service also seems to be regarded as a decisive factor of product attractiveness of wedding venue p_H. As described above, the superiority factor analysis enables us to capture the strength of the given product against a specific competitor in the form of the list of keywords.

5 EVALUATION EXPERIMENT

We evaluate our proposed method as useful for e-commerce marketing by comparing user perceptions, as estimated using mining log data, with actual perceptions investigated by conducting a user survey. Our objective is to extract user perception of the superiority relation in product attractiveness and its decisive factors in an e-commerce website. We conducted a user survey of couples who used Zexy for the reservation for venue tours and who held a ceremony at one of the candidate venues for investigating actual user perception, which are answered explicitly through the survey.

We conduct two experiments in this section. The first experiment is evaluation of the correlation between the product networks: one is extracted from the user survey results. The other is calculated from the log data. We infer that the log data can be a good alternative of survey results in terms of extracting products' competitive relations and win–lose relations, which have been defined in this paper. The other experiment evaluates the number of matches of factor words. We presume a baseline method that outputs factor words using only winner product review texts and compare the number of matches to the actual factor words extracted from the user survey results.

[2]Mecab http://taku910.github.io/mecab/

Table 3: Correspondence between the user survey results D and log data $\hat{D}$.

	User survey D	Log data $\hat{D}$
Users	Responders (N=173)	Visitors (N=202)
Browsing	Attending tour	Browsing
Purchasing	Holding a ceremony	Making a reservation
Text	Reason for selection	Product review

5.1 Experimental Setup

In both experiments, we use Zexy as the objective website and specifically examine the 10 wedding venues selected in Section 4. The user survey was administered from January 23, 2012 to December 14, 2013, targeting the respondents who booked multiple wedding venue tours on Zexy and who finally held a ceremony by selecting one of them. The survey asks respondents about the wedding venue at which they held the wedding ceremony and the reason they ultimately selected that wedding venue as a result of that comparison. The user survey results can be regarded as representing user preferences for wedding venue selection because they are explicit responses by users. Evaluating the methods using multiple datasets is preferred, but we specifically examine this website in this paper because it is rare to have a large-scale e-commerce website accompanied by an offline user survey that studies actual users' preferences.

Generally speaking, the user survey setup entails huge costs exemplified by questionnaire distribution and data collection. However, the log data are recorded implicitly by observing user activities, which does not impose any burden on users to input information manually. Therefore, estimating the product relation from the log data instead of conducting actual user surveys is valuable. As described in this paper, we designate the product relation extracted from the user survey results as the *actual product relation* and the factor words as *actual factor words*. We evaluate whether we can estimate them, or not, by application of our proposed methods to the log data.

Table 3 presents correspondence between the user survey results and the log data. We use the sessions from log data showing both browsing and purchasing behavior for the target 10 wedding venues, i.e., the sessions contribute to form a win–lose relation among them. The browsing behavior on the website can correspond to the offline behavior to attend the wedding venue tours. Making reservations for venue tours on the website can correspond to the offline behavior to decide the wedding venue at which to hold the wedding ceremony eventually. In addition, the review texts can correspond to survey responses explaining why they selected the wedding venue. According to these correspondences, we can extract the product relation by application of our proposed method on the user survey results as well.

5.2 Evaluation of Product Relation Analysis

First, we evaluate the correlation of the competitive relations. We calculate the estimated weight $\hat{w}^C$ from the log data $\hat{D}$ and the actual weight w^C from the user survey results D for 45 competitive relations between the given 10 venues. Hereinafter, we assume that the relation between the two products is determined independently of other products' relations. Results revealed a Pearson correlation coefficient of 0.685 with a p-value of 2.06×10^{-7}, which is a significant correlation. Actually, 6 out of 45 pairs had a value of zero in either the actual weight or the estimated weight, but the remainder of them had positive weights in both. The results demonstrated that our proposed method yields good estimation of the competitive relation among the products.

Second, we evaluate the correlation of the win–lose relations. We can extract 90 win–lose relations from the 10 venues because the win–lose relations have directions unlike the competitive relations. Along with the competitive relation analysis, we assume that the relations between the two products are determined independently from others. We calculated the estimated weight $\hat{w}^{WL}$ from the log data $\hat{D}$ and calculated the actual weight w^{WL} from the user survey results D. Results show that the Pearson correlation coefficient is 0.648, with a p-value of 5.02×10^{-12}, which indicates a significant correlation between them. Actually, 20 out of 90 pairs had a zero value in either the actual weight or the estimated weight, but the remainder of them had positive weights in both. Results show that our proposed method is making a good estimation of the win–lose relation among the products.

5.3 Evaluation of Superiority Factor Analysis

Finally, we evaluate the performance of the superiority factor analysis. To begin with, we explain the extraction method of the actual factor words of wedding venue p_i against wedding venue p_j from the user survey results. We select the responses of which responder selected wedding venue p_i to hold the wedding ceremony after attending a wedding venue tour of wedding venue p_j, i.e., the responses which perceive product p_i as the winner and product p_j as the loser. We aggregate their reason for selection into one text and conduct morphological analysis of it to process it into a set of noun phrases. We assume this set as the actual factor words $T_{p_i \succ p_j}$.

We calculate the estimated factor words using our proposed method $T^{proposed}_{p_i \succ p_j}$ by following the same procedure introduced in Section 3, which extracts top K keywords by taking the product reviews of all wedding venues in Tokyo as the corpus. For evaluation of the effectiveness of our proposed method, we set a baseline method to compare which does not consider any win–lose relation. The baseline method assumes the keywords extracted from the product review of the winner product p_i as the estimated factor words $T^{baseline}_{p_i}$, no matter which the loser product is. In other words, it simply assumes that the characteristics of the winner product is the decisive factors of the win–lose relation for any loser product. This method also uses the same corpus as the proposed method and returns the top K words with the highest tf–idf score.

This experiment evaluates the performance of the analysis methods using the number of matches between the actual factor words $T_{p_i \succ p_j}$ and estimated factor words using the proposed method $T^{proposed}_{p_i \succ p_j}$ and the baseline method $T^{baseline}_{p_i}$ for the win–lose pair (p_i, p_j). We use these metrics assuming that the actual factor words are capturing the aspects and characteristics users care most. We set $K = 20$ for this experiment. Additionally, we skip evaluation of the pair if the number of actual factor words is less than 5 or if

Table 4: Factor words of product p_D against product p_G. Matched words are highlighted in bold.

Method	Factor Words
Baseline $T^{baseline}_{p_D}$	map, cathedral, forbidden, she, Akka, order, **impression**, problem, standard, cloud, **stained glass**, **church**, European, minute, overall, exchange, movie, ring, Omotesando, bringing (3 matches)
Proposed $T^{proposed}_{p_D \succ p_G}$	**chapel**, **hospitality**, **ceremony**, guest, feeling, day, **impression**, **staff**, **stained glass**, **banquet**, dish, **atmosphere**, **church**, location, **San**, **photograph**, **lovely**, **venue**, Omotesando, weddings (13 matches)

the number of estimated factor words does not meet K by either method.

Results show that 45 out of 90 pairs satisfied the criteria described above. Among them, 23 pairs showed that our proposed method outperforms the baseline method. In all, 21 pairs showed a tie. The remaining one pair showed that the baseline method outperformed our proposed method. In all, 1548 actual factor words were extracted and 284 estimated factor words matched our proposed method, whereas 230 factor words matched with the baseline method. As a result of the chi-square test of independence, we found a statistically significant difference between the match rates of the two groups with the p-value of 3.09%. Therefore, our proposed method is better than the baseline method at estimating the actual factor words from log data.

Table 4 presents the estimated factor words of wedding venue p_D against wedding venue p_G, which demonstrates that our proposed method succeeded in estimating the actual factor words with the greatest discrepancy in terms of the match ratio. Against 68 actual factor words, our proposed method estimated 13 words, whereas the baseline method matched 3 words. Wedding venue p_D knows to offer Western style weddings. The baseline method can capture that feature with keywords such as *church* and *stained glass*. However, this is a well-known point of appeal with this wedding venue, which is not such an interesting finding. In contrast, our proposed method captured another aspect of the characteristics such as *hospitality*, *atmosphere*, and *staff*. By reviewing these words, we ascertain that their customer services and hospitality are also evaluated by customers when it is compared to wedding venue p_G.

Experimental results show that the baseline method can extract features of the purchased product, but it cannot extract features about which users care in a broader sense. Sometimes, it becomes too specific to the product and sometimes becomes too general to explain the decisive factor. Our proposed method, however, can extract the purchased product's specific characteristics and can extract important features that satisfy users for whom preferences support the given win–lose relation. Our proposed method can offer appropriate factor words by widening the scope of consideration from the purchased product to users' preferences, which contribute to formation of the product relations.

6 DISCUSSION

We compared the product relations calculated from the log data and the user survey results, which presumably reflects the actual perceived superiority relation among products. Results show that their competitive and win–lose relations are mutually correlated. These results show that our proposed method can capture the actual users' perceptions by analyzing log data in the form of product relations defined in this paper. Therefore, these results also show that log data can be a good alternative of the user survey. For that reason, we might save the cost for conducting user surveys, which entails expensive processing work such as questionnaire distribution and collection, and data inputs.

We also evaluated the superiority factor analysis method using the actual dataset. Experimental results show that our proposed method provides good estimation of the actual factor words extracted from the user survey results than the baseline method which specifically examines the purchased products' review text and which does not consider win–lose relations of products. Therefore, our proposed method can extract the features perceived as important by patrons from product reviews by considering win–lose relations among the products.

Our proposed method specifically examines the keywords and does not capture the sentiment on features, which might represent negative opinions on the factor words. Nevertheless, extracting factor words is useful to capture aspects of the product or service that customers perceive as important. It will be a helpful hint when marketers dig into the product reviews closely. We used a combination of simple approaches aiming to validate the existence of factor words, but the possibility exists that we can conduct more sophisticated analyses that consider the polarity of the sentence. Additionally, the keyword extraction may be improved by considering the reviews related to $P_{p_i \succ p_j}$ in comparison to ones related to $P_{p_j \succ p_i}$.

If the e-commerce website is handling multiple categories of products, then we might require some technique to build product networks within each category to avoid extracting meaningless cross-category relations such as "camera versus detergent". Some e-commerce websites such as consumer-to-consumer commerce and online auctions might have multiple sellers offering different prices for one product. In this case, the attractiveness of each offering can be included as a noise factor in the product network. Our proposed method is a generalized method that is compatible with any e-commerce website as long as log data are available, but it will work better with marketplaces for which the law of one price holds.

Additionally, it is noteworthy that competitive and win–lose relations are defined between substitute products, not between complementary products. This assumption holds on Zexy because it is a wedding portal website. All handled products are wedding venues, which are inherently competitors. However, generic e-commerce websites such as Amazon[3] handle complementary products within a category (e.g., a laptop and its battery in the electronics category). Our proposed method is directly applicable to some e-commerce websites that are specialized for specific products, but some technique might be necessary to identify and filter complementary

[3] Amazon https://www.amazon.com/

products when we apply this method to a generic e-commerce website.

As seen in some win–lose relations such as between wedding venues p_A, p_C, and p_E in Figure 5, there can be arrows of comparable thickness going in either direction. We can regard them as cancelled, but it also means that a comparable number of users perceive the win–lose relation in either direction. If we can capture user clusters that separate the conflicting win–lose relation well, then the information will be useful to capture multiple dynamics of product competition, which differ depending on user characteristics. This information is also useful for product marketers to elucidate the characteristics of their already convinced customers and those to whom they must appeal.

The possibility exists that the promoted products might tend to be the winner product during that period if there is a promotional campaign on the objective e-commerce website. However, if users visit to the promoted items directly without considering other products, then the behavior does not affect the original win–lose relation of products because there is no loser product in such a case. Similarly, user might purchase a product in the e-commerce website because they offer a lower price than their competitors aside from product attractiveness. In that case, it is likely that users have already decided the item to purchase in their mind when they visit the website. Therefore, other products are less likely to be browsed, which means there is little effect on the product network. Our proposed method has robust characteristics to temporal events such as promotional marketing campaigns and the difference of selling price across e-commerce websites.

For cases in which numerous products are handled in the website, we can think of some techniques to remove products that are not purchased much or merge similar products. These techniques will be useful to avoid making the product network sparse. For cases in which very few products are handled, the possibility exists that the product network becomes too dense and difficult for readers to grasp an overview. We can think of a technique to filter edges with small weight.

If there are too many users who belong to the analysis, then it is possible to extract various product relations, but the possibility exists that various user preferences are mixed to the product relation, which makes it difficult to analyze their underlying superiority factors. In this case, we might be able to categorize the users beforehand based on their attributes and be able to conduct product relation analysis for each user category to extract more sophisticated and actionable insights. However, our proposed method functions well even if there are few users because it can scale up the data volume using users' browsing behavior in addition to purchasing behavior.

7 CONCLUSION

We proposed a new product relation analysis method that uses the log data of e-commerce websites to reveal the superiority relation in product attractiveness. Our proposed method also analyzes the superiority factor using text data associated with the products. We applied our analytical method to a Japanese wedding portal website, Zexy, to extract the product network and superiority factors from their actual dataset. In the evaluation experiment, we compared product networks extracted from the website log data and user survey results. Experimental results showed that the weights of both product networks are correlated with statistical confidence, which demonstrates that the log data can be a good alternative of the user survey for extracting user preferences. Results show that our proposed superiority factor analysis method can estimate the actual decisive factors of product attractiveness, which is extracted from user survey results. Our proposed method is beneficial for marketers and product managers to understand the competitive advantages of a given product and consolidate their product differentiation strategies.

REFERENCES

[1] Nikolay Archak, Anindya Ghose, and Panagiotis G Ipeirotis. 2007. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 56–65.

[2] Ming C Hao, Umeshwar Dayal, Meichun Hsu, Thomas Sprenger, and Markus H Gross. 2001. *Visualization of directed associations in e-commerce transaction data*. Springer, Vienna.

[3] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 168–177.

[4] Mathieu Jacomy, Sebastien Heymann, Tommaso Venturini, and Mathieu Bastian. 2011. *Forceatlas2, a continuous graph layout algorithm for handy network visualization.* Medialab Center of Research.

[5] Theodoros Lappas, Mark Crovella, and Evimaria Terzi. 2012. Selecting a characteristic set of reviews. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 832–840.

[6] James M Lattin and Leigh McAlister. 1985. Using a variety-seeking model to identify substitute and complementary relationships among competing products. *Journal of Marketing Research* 22, 3 (1985), 330–339.

[7] Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015. Inferring networks of substitutable and complementary products. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 785–794.

[8] Troy Raeder and Nitesh V Chawla. 2009. Modeling a store's product space as a social network. In *Proceedings of International Conference on Advances in Social Network Analysis and Mining*. IEEE Computer Society, Washington, DC, USA, 164–169.

[9] Troy Raeder and Nitesh V Chawla. 2011. Market basket analysis with networks. *Social Network Analysis and Mining* 1, 2 (2011), 97–113.

[10] Allan D Shocker, Barry L Bayus, and Namwoon Kim. 2004. Product complements and substitutes in the real world: the relevance of other products. *Journal of Marketing* 68, 1 (2004), 28–40.

[11] Zhiang Wu, Youquan Wang, Yaqiong Wang, Junjie Wu, Jie Cao, and Lu Zhang. 2015. Spammers detection from product reviews: a hybrid model. In *Proceedings of the IEEE International Conference on Data Mining*. IEEE Computer Society, Washington, DC, USA, 1039–1044.

[12] Jiao Xu, Chris Forman, Jun B Kim, and Koert Van Ittersum. 2014. News media channels: complements or substitutes? Evidence from mobile phone usage. *Journal of Marketing* 78, 4 (2014), 97–112.

[13] Jiaqian Zheng, Xiaoyuan Wu, Junyu Niu, and Alvaro Bolivar. 2009. Substitutes or complements: another step forward in recommendations. In *Proceedings of the Tenth ACM Conference on Electronic Commerce*. ACM, New York, NY, USA, 139–146.

[14] Dmitry Zinoviev, Zhen Zhu, and Kate Li. 2015. Building mini-categories in product networks. In *Complex Networks VI*. Vol. 597. Springer, Cham, 179–190.