# Predicting User's Online Shopping Tendency During Shopping Holidays

Cheng-You Lien, Guo-Jhen Bai, Ting-Rui Chen, Hung-Hsuan Chen
Computer Science and Information Engineering
National Central University
{littlelienpeanut, ivy2350442}@gmail.com, ray941216@g.ncu.edu.tw, hhchen@ncu.edu.tw

***Abstract*—The number of sales during the shopping holidays continues growing in recent years. Thus, many E-Commerce (EC) websites spend much money and effort for marketing before these shopping holidays. However, in this study we found that only part of the Internet users indeed visited the EC-websites more often than usual during the shopping holidays. Thus, the increase of the sales probably comes from few individuals. Additionally, we found that users' tendency to visit the EC websites during the shopping holiday is predictable based on simple supervised classifiers. Thus, an EC website runner can identify the potential visitors and non-visitors beforehand and apply different marketing strategy to different users.**

## I. Introduction

The sales volume of the e-commerce websites during the shopping holidays (e.g., the Singles' Day, Christmas, and Moon Festival) continues growing strongly in recent years [1], [2]. This seems to suggest that users tend to visit the EC websites more often than usual during these periods. However, we found that only part of the users indeed visited the EC-websites more frequently during the shopping holidays, based on our collected user logs. Thus, the increase of the sales probably comes from few individuals.

This observation motivates the study: can we predict the potential visitors and non-visitors of EC websites during the specified shopping holiday? If we can classify the potential visitors from the non-visitors, we can probably apply different marketing strategies to different groups of users. For example, if a user is highly likely to visit shopping websites during the shopping holidays, a EC website runner should probably show more advertisements or special offers to this individual, so that she/he will be more likely to visit the owner's website instead of visiting the competitors'.

This paper validates that it is possible to predict the potential visitors and non-visitors of EC websites during the shopping holidays. Specifically, we make the following contributions. First, we compared users' usual browsing behavior and during the shopping holidays. We found that the daily average ratio of the visits on the shopping websites stays steadily between $4\%$ and $10\%$. Most users do not increase their visits to the shopping websites during the shopping holidays. This is very different from the usual belief. Thus, the increase of the sales during the shopping holidays probably comes from few individuals. Second, based on simple supervised learning algorithms with appropriate features, such as their browsing history and demographic information, we can predict the users who may visit the shopping websites more frequently during the shopping holidays. The online retailers may use such a model to adjust their marketing strategies, e.g., sending the coupons to the users who need more incentive for shopping. Last but not the least, we showed the statistics of users' cross-website browsing logs. Such a dataset is usually owned by large portal sites (e.g., Google) or large ISPs (e.g., Comcast) and normally kept in secret. We plan to release an anonymized version of the dataset.

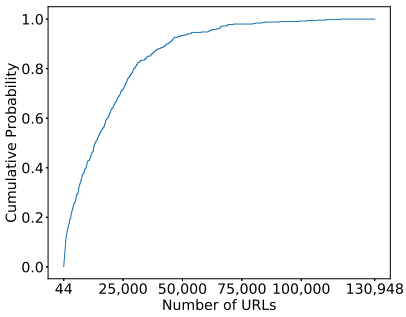



Fig. 1: The empirical cumulative density function of the number of visited URLs per user.

The rest of the paper is organized as follows. We will introduce the experiment dataset in Section II. Section III reports the experiments of shopping tendency prediction for the shopping holidays. We review related work in Section IV. Finally, we discuss the limitations of the study and future work in Section V.

## II. Description of the Data

### A. Summary of the data

TABLE I: A statistical summary of the number of visited URLs per user.

| min | 1st Quartile | Median | Mean | 3rd Quartile | max |
|---|---|---|---|---|---|
| 44 | 4,239 | 13,335 | 19,103 | 26,698 | 130,992 |

We recruited 517 individuals as the target user for the study. Specifically, we collected these users' complete browsing history stored in their Google Chrome browsers. Most of these browsing histories were recorded from August 2016 to

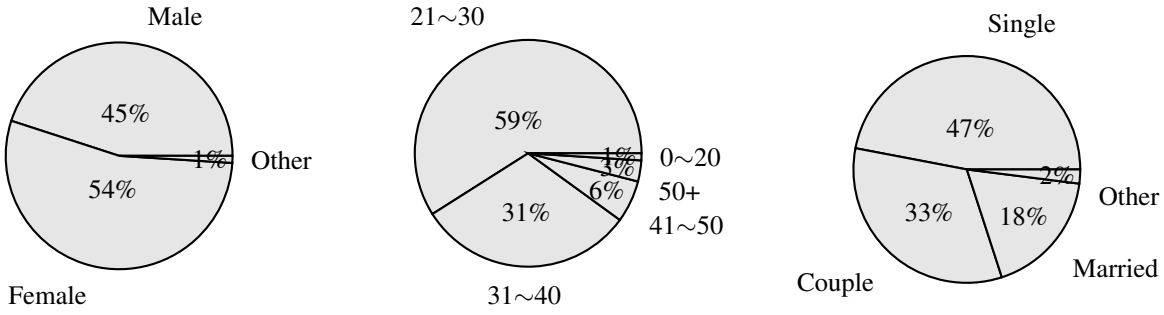



Fig. 2: The pie chart of: (a) Gender (b) Age (c) Relationship status

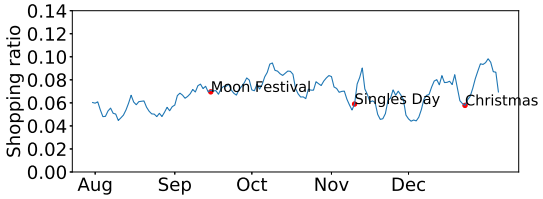



Fig. 3: Users' visiting ratio on shopping websites.

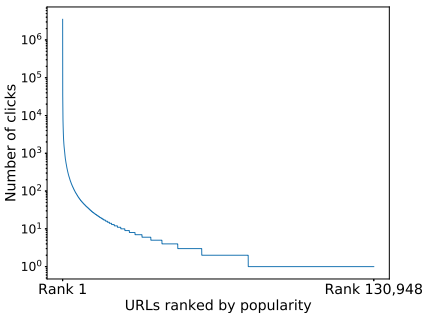



Fig. 4: The number of clicks of the websites ranked by popularity

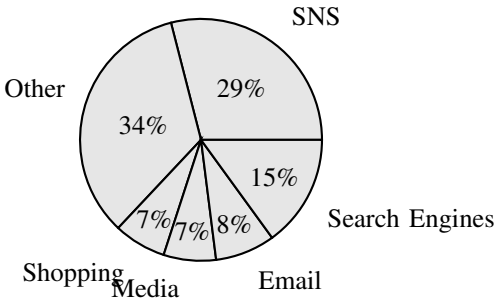



Fig. 5: The proportions of the visited website categories

December 2016. All these individuals reported that they are familiar with Internet and have experience in online shopping.

Although the number of users is not much, we have a very detailed record of each user's browsing history: these 517 users totally contribute 12,653,625 browsing records. As a result, we can track these user's cross-site footprints thoroughly. A statistical summary and an empirical cumulative density function of the number on the visited URLs per user are shown in Table I and Figure 1 respectively.

Additionally, we ask each of these 517 users to fill a questionnaire so we obtained these users' self-disclosed demographic information, such as the age, gender, and relationship status (single, couple, married, and others). Figure 2 shows the pie chart of these demographic information.

To visualize users' habits on visiting the shopping websites, we show users' ratio of visits on the shopping websites on continuous dates. Specifically, for each user, we define the shopping ratio on a particular date as the number of visits on the shopping websites divided by the total visits. The average shopping ratio on each day is shown in Figure 3. As demonstrated, the shopping ratios stay between $4\%$ and $10\%$ through the study period. In addition, even during the three large shopping holidays (Moon Festival, Singles' Day, and Christmas), the shopping ratios do not show significant rising. Therefore, the increase of the sales during the shopping holidays may come from few individuals. Correctly predicting the users who may visit the shopping websites more frequently during these periods may bring a huge benefit to the EC website runners.

### B. Data preprocessing

The original dataset contains all these users' entire browsing logs in their Chrome browsers. We found that users' visited websites are highly imbalanced. For example, $27.8\%$ of the users visited the most popular website facebook.com, but most websites are visited by few (even single) individuals. Figure 4 shows the number of clicks of the websites sorted by popularity ($y$-axis in log scale). As a result, the skewness of the distribution of the URLs affect the performance of the classifiers: the popular websites have little discriminative power, since many user frequently visits these websites. On the other hand, although the uncommon websites could be very powerful in discriminating certain characteristics of the

users, we can only adopt the feature on limited users, since most users do not visit the uncommon websites.

To deal with the skewed distribution of the websites, we converted all the URLs to 73 categories based on Web Filter Lookup[1], a public service to convert URLs into categories. For example, https://www.google.com/ is converted to "Search Engines and Portals", and https://www.facebook.com/ is converted as "Social Networking". For each user, we compute her/his accumulated visits of each category on each day. Eventually, we found that Social Networking Services (SNS) and Search Engines are the most popular categories, followed by Email, Shopping, and Media. Figure 5 shows the proportion of the categories in the browsing logs.

## III. Shopping tendency prediction

### A. Experiment setup

We selected three shopping holidays during August 2016 and December 2016, because most of our available browsing histories are within this period. The three selected holidays are the Moon Festival (9/15/2016), the Single Day (11/11/2016), and Christmas (12/25/2016). We manually examined the Internet Archive[2] and several web caches to ensure many popular EC websites (e.g., PChome, Momo, GoHappy) in Taiwan provided special promotions on selected items during these shopping holidays.

We define a user as a positive instance if the user's average shopping ratio increases during the target shopping holiday. We set the period of a shopping holiday to be the date of the holiday plus 6 days before it. For example, the Moon festival shopping holiday is 9/9/2016 - 9/15/2016 (the date of Moon festival).

### B. Training models and features

We utilized several popular supervised classifiers, including K-nearest neighbors (KNN), support vector machine (SVM), random forest (RF), and logistic regression (LR), to make predictions. We selected the hyper-parameters based on grid searches. Eventually, we set the hyper-parameter $k$ as 7 for KNN; we set $C$ the inverse of the regularization strength to $2.0$ for logistic regression and $1.0$ for SVM (with RBF kernel); and the number of trees to $10$ for random forest.

We selected features from users' demographical information and users' browsing history. The demographical features include users' genders, ages, and relationship status (single, couple, married, and others). Each user's browsing-related information is compiled from the three website categories with the highest proportions during her/his normal periods (i.e., before the start of the shopping holidays).

### C. Performance of the prediction

We randomly split the users into training and the test groups for 20 times. The average of the training and the test AUCs of the ROC curves are shown in Table II. It appears that SVM and Logistic Regression perform better than the rest. In addition, as the number of training dates becomes larger, the test AUCs become larger. Figure 6 shows the ROC curves for the three holidays from one of the 20 trials.

TABLE II: The average AUCs of the training and the test datasets of different classifiers on the three holidays

| | Moon Festival | | Singles Day | | Christmas | |
|---|---|---|---|---|---|---|
| | Training | Test | Training | Test | Training | Test |
| KNN | 0.71 | 0.55 | 0.75 | 0.63 | 0.78 | 0.72 |
| LR | 0.73 | **0.65** | 0.70 | 0.61 | 0.82 | 0.73 |
| SVM | 0.74 | 0.64 | 0.84 | **0.64** | 0.84 | **0.77** |
| RF | 0.99 | 0.60 | 0.99 | 0.60 | 0.99 | 0.68 |

### D. Training size vs performance

In Table II, we observed that the training AUCs are consistently larger than the test AUCs. This suggests that the classifiers *overfit* the training data. Thus, we surmise that, if we could obtain the logs from a larger group of users, it is possible to improve the performance of the predictions.

To validate the idea, we compiled three (training, test) pairs based on all the available datasets. Specifically, we set the size of the training data to $50\%$, $70\%$, and $90\%$ of all the available datasets, and treat the rest users as the test data for validation.

Figure 7 shows the test AUCs of the three different settings. As shown, the test AUCs indeed increase dramatically as the training data increases from $50\%$ to $90\%$. The curve trend of the test AUCs continues growing. Therefore, it is very likely that the performance of the prediction can further be improved, once more training instances are available.

## IV. Related Work

Recommendation algorithms are typically categorized into four types: content-based filtering [3], collaborative filtering [4], [5], context-aware [6], and hybrid methods. These algorithms rely on users' features, items' features, and users' online behaviors on the items to make recommendations. However, all these methods require users to visit the EC website so that the recommendations have chances to appear on users' browsers.

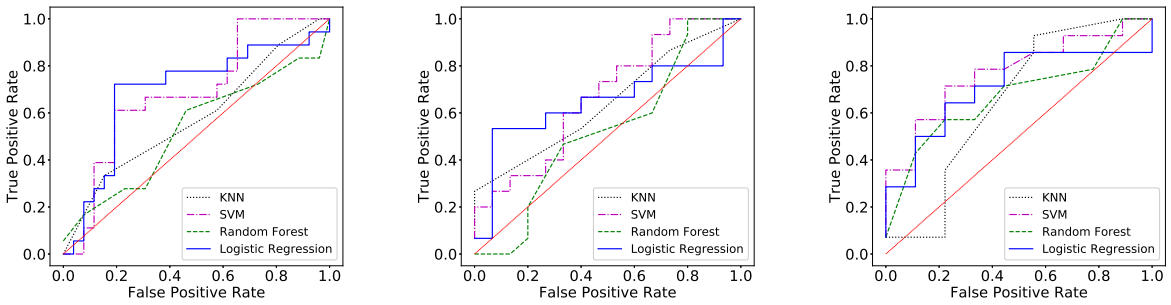
To recommend items in a more aggressive manner, the electronic direct mails (EDMs) could be included in the marketing mix. Such a method sends the recommended items or special offers to the users' emails directly. However, studies have shown that the recipients of the unsolicited EDMs sometimes feel intrusive [7]. Thus, it is essential to send the EDMs (and customized EDMs) only to the individuals who might be interested in the items included in the EDMs [8].

Additionally, some argue that the EDMs or the advertisements should also be sent to the users who have strong influence to affect other users' decisions. Many studies proposed methods to discover the key persons on the Internet based on network analysis [9]–[11], and hope to increase the sales volume through the word-of-mouth marketing.

This paper suggests to send the advertisements to the customers from a different perspective. We predict which users

[1]http://www.fortiguard.com/webfilter

[2]https://archive.org/web/




(a) The Moon Festival (9/15) (b) The Singles Day (11/11) (c) Christmas (12/25)

Fig. 6: The ROC curves of the three holidays based on the test data

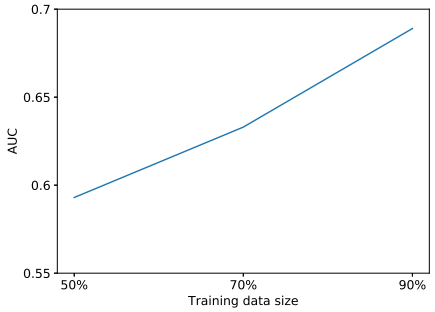



Fig. 7: The test AUCs when using 50%, 70%, and 90% of the available data instances as the training data (using logistic regression).

are likely to visit the EC websites more frequently during the shopping holidays. The website runners can apply different advertisement strategies to different groups of users based on the predicted results.

## V. Discussion and Future Work

Although most EC websites sell more items during the shopping holidays, this paper found that such increases may come from few individuals. Thus, if an EC website runner can identify the potential visitors beforehand, it may have more advantage in creating customized marketing strategies toward different types of users. This paper shows that indeed users' tendency to visit the EC websites during shopping holidays is predictable by simple supervised classifiers. In addition, we show that the prediction performance can further be improved once we acquire more training data.

The main limitation of the paper is the user size for experiment. One of our future work is collecting the data from more users. A larger dataset will make the study more credible and possibly improve the prediction performance. The other limitation is that we can only collect users' visited pages, but visiting a shopping website more frequently does not necessarily mean more purchases. Unfortunately, the purchasing information requires server-side authentication, which is difficult to obtain in practice.

For future work, we plan to release the anonymized log for public research and collecting more data. In addition to the shopping holidays, we are also interested in investigating how other life events may influence users' online behaviors.

## VI. Acknowledgement

We appreciate financial support from the Ministry of Science Technology (MOST 105-2218-E-008-015) and the Industrial Technology Research Institute (ITRI 106-W100-21A2). We are grateful to the National Center for High-performance Computing (NCHC) for computer time and facilities.

## References

[1] "E-Commerce Continues To Be The Bright Spot For Holiday Sales," https://www.forbes.com/sites/shoptalk/2016/12/27/ecommerce-continues-to-be-the-bright-spot-for-holiday-sales/#28b1f6bb2780, accessed: 2017-06-15.

[2] "Alibaba's Singles' Day: What We Know About The World's Biggest Shopping Event," https://www.forbes.com/sites/franklavin/2016/11/06/alibabas-singles-day-what-we-know-about-the-worlds-biggest-shopping-event/#687e1e636da7, accessed: 2017-06-15.

[3] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The adaptive web*. Springer, 2007, pp. 325–341.

[4] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[5] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet computing*, vol. 7, no. 1, pp. 76–80, 2003.

[6] G. Adomavicius and A. Tuzhilin, "Context-aware recommender systems," in *Recommender systems handbook*. Springer, 2011, pp. 217–253.

[7] M. Morimoto and S. Chang, "Consumers attitudes toward unsolicited commercial e-mail and postal direct mail marketing methods: intrusiveness, perceived loss of control, and irritation," *Journal of Interactive Advertising*, vol. 7, no. 1, pp. 1–11, 2006.

[8] D. J. Xu, S. S. Liao, and Q. Li, "Combining empirical experimentation and modeling techniques: A design research approach for personalized mobile advertising applications," *Decision support systems*, vol. 44, no. 3, pp. 710–724, 2008.

[9] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 57–66.

[10] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[11] J. Leskovec, A. Singh, and J. Kleinberg, "Patterns of influence in a recommendation network," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2006, pp. 380–389.