Mobile Networks and Applications manuscript No.
(will be inserted by the editor)

Profiling Users by Online Shopping Behaviors

Huan Yan · Zifeng Wang · Tzu-Heng Lin · Yong Li · Depeng Jin

Received: date / Accepted: date

Abstract Online shopping has been prevalent in our daily life. Profiling users and understanding their browsing behaviors are critical for enhancing shopping experience and maximizing sales revenue. In this paper, based on a one-month dataset recording 2 million users' 67 million online shopping and browsing logs, we seek to understand how users browse and shop products, and how distinct these behaviors are. We find that there exist dedicate groups of users that prefer certain product categories corresponding to similar demands. Moreover, distinct differences of behaviors exist in categories, where repetitive and targeted browsing are two major prevalent patterns.

Keywords

H. Yan
Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

ZF. Wang
Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

TH. Lin
Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

Y. Li (corresponding author)
Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China
E-mail: liyong07@tsinghua.edu.cn

DP. Jin
Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China

1 Introduction

The rapid growth of Internet usage drives the popularity of online shopping. Especially, online shopping demands can be boosted along with advanced vehicular telematics over heterogeneous wireless networks [1], and large amounts of real-time online data can also be collected by using distributed information estimation technologies [2].According to the latest eMarketer's forecast, it will increase 5.2% from 2015 to 2019, and account for 12.4% of global retail sales by 2019 [3]. To attract more users and maximize the revenue, e-commerce business (e.g., *Amazon*) strives to provide better services, *i.e.*, designing personalized recommendation systems to improve users' shopping experience [5–7].

One of the fundamental problems here is to thoroughly understand users' purchasing demands and shopping patterns [11, 12]. Compared to traditional survey questionnaire, data-driven behavior analysis can comprehensively reveal what users prefer and how they select products [13, 16–19]. Previous work studied repeated consumption behavior [4,8], but how users browse different kinds of products online, and how distinct their behaviors are, are still unknown.The work of heterogeneous systems from Qiu et al. [14] has proposed the way to reduce cost of complicate heterogeneous data and system. Also, their work on online system [15] and green cloud [10] had provide valuable guidance in online data processing and cost reduction.

To investigate this problem, we use a large-scale dataset that contains user online shopping and browsing logs at one of major e-commerce businesses in China. Our dataset is collected from one of major ISPs, which contains 67 million browsing records of 2 million users in Shanghai from March 2 to March 31 in 2015. First, we use a co-clustering method to cluster both users and categories of browsed products simultaneously. Then, we seek to understand the characteristic of shopping behaviors based on average consecu-

tive browses. We obtain two major findings, summarized as follows:

– There are both homogeneous (e.g., users browsing one category) and heterogeneous (e.g., users browsing diverse categories) groups of users.
– There exist distinct differences of user browsing behaviors in different categories. Repetitive and targeted browsing are two prevalently recognized patterns.

Our findings are useful in designing customized online shopping web systems for dedicated groups of users by adapting to their personal consumption behaviors. In addition, from the perspective of ISPs, they can characterize the user profile that has potential commercial value.

2 Dataset

Our dataset is collected through deep packet inspection appliances at the gateways of ISP, which contains complete shopping and browsing logs of a large online e-commerce platform. Each entry in the logs is characterized by user ID, timestamp and requested URL. To obtain detailed information, we crawl URLs at the e-commerce website and obtain the corresponding product category of each browsing request.

In summary, we obtain more than 67 million browsing requests of 2,141,951 users who browse over 15 million products, which are classified into 28 categories (e.g., Clothing, Books, Phone & Accessories). We plot the distribution of the number of products, users and browses in 11 major categories[1] that occupy 82.35% of browses, as shown in Figure 1. Although *phone & accessories* attracts most browses, the number of products belonging to this category is relatively small. In contrast, *(E-)Books & CDs* owns a large amount of products but relatively low browses. The reason is that user demands are different in shopping products of diverse categories, and they exhibit different shopping behaviors when browsing different kinds of products.

3 Metric and Methodology

With the goal of profiling users by their online shopping and browsing logs, we now describe our metrics and methodology. We denote $\Omega = \{c_k\}(1 \leq k \leq 28)$ as the set of product categories and W_k as the number of products belonging to c_k. For a given user i, we model a user's browsing record as a sequence of browsing events s_{i_j}: $S_i = \{s_{i_1}, \Delta t_{i_1}, s_{i_2}, \Delta t_{i_2}, ..., s_{i_j}, \Delta t_{i_j}, ...\}$ with Δt_{i_j} representing the time gap between two consecutive events s_{i_j} and $s_{i_{j+1}}$. We denote $M_i^k = \sum_{j=1}^{N_i} I_{c_k}(s_{i_j})$ as the total number of browsing in k-th category by user i, where N_i is the number of browses by user i and $I_A(x)$ is the indicator function.

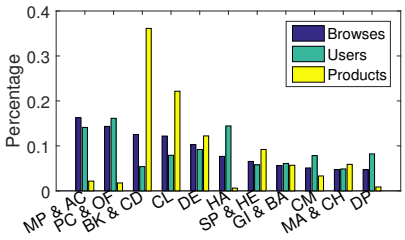

Fig. 1: Distribution of the number of browses, users and products in major 11 categories[1].

– **Browsing Entropy**: It measures how diversely users browse the products of different categories, defined as

$$E_i = \frac{-\sum_{k=1}^{28} \frac{M_i^k}{\sum_{k=1}^{28} M_i^k} \log_2 \frac{M_i^k}{\sum_{k=1}^{28} M_i^k}}{\log_2 \sum_{k=1}^{28} I_{\{>0\}}(M_i^k)}. \quad (1)$$

Its value ranges from 0 to 1. A higher value indicates more uniformly distribution among all categories. If the user only browses one category, it is 0.
– **Repetitive Ratio**: This measures how frequently products of the same category are browsed by users, expressed as $R(k) = \sum_i M_i^k / W_k$, where a higher value indicates that users more frequently browse products of k-th category.
– **Co-Clustering of Users and Categories**: Since users and their browsing categories are associated with each other, we need to cluster both of them simultaneously. We use *Phantom* [9] to perform divisive hierarchical co-clustering. We calculate the normalized number of browses in each category per user, then obtain a feature matrix with users on each row and categories on each column, which is the input of the co-clustering algorithm.
To evaluate the effectiveness of co-clustering, we define the average distance to each cluster as follows:

$$D(m,k) = \frac{1}{p_m} \sum_{i=1}^{i=p_m} |(\mathbf{F}_m^i - \bar{\mathbf{F}}_k)|, \quad (2)$$

where $\bar{\mathbf{F}}_k = \frac{1}{p_k} \sum_{i=1}^{i=p_k} \mathbf{F}_k^i$. $\mathbf{F}_m^i$ represents the array consisting of the normalized browses in each category by user i in cluster m, and p_m denotes the number of users in cluster m. In particular, if $D(m,m) < D(m,k)$ $(m \neq k)$ is satisfied, users in cluster m have higher similarity compared with that in other clusters.

[1] These include Phone & Accessories (MP & AC), PC & Office (PC & OF), Books & CDs (BK & CD), Clothes (CL), House Decorations (DE), Household Appliances (HA), Sports & Health (SP & HE), Gifts & Bags (GI & BA), Cosmetics (CM), Maternity & Child (MA & CH) and Digital Products (DP).

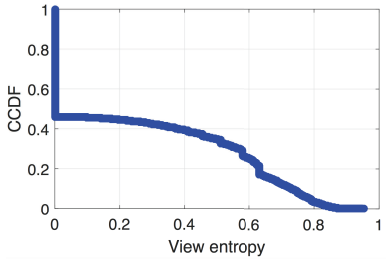

Fig. 2: Distribution of browsing entropy that is defined in (1).

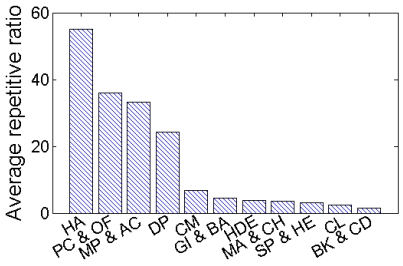

Fig. 3: Average browsing repetitive ratio in 11 major categories[1].

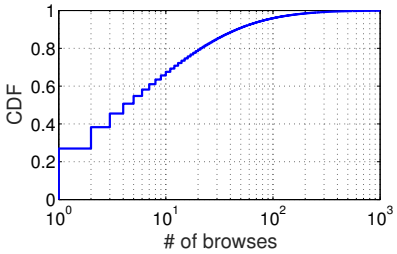

Fig. 4: Distribution of the number of browsing per user: 27% of users have only one browse, and 96% of users have less than 100.

– **Category-based Browsing Behavior Analysis**: In order to reveal how distinct users browse different categories, we partition the browsing sequence S_i into different sessions by a time threshold, exceeding which indicates a user is offline. Then, we count the number of consecutive browses on each category in each session within S_i. Finally, we average the consecutive browsing on k-th category by user i.

4 Results

In this section, we leverage the above metrics to analyze the online shopping behaviors based on our collected dataset, which is described in details in Section 2.

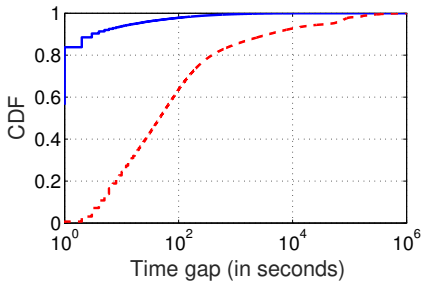

Fig. 5: Distribution of the time gaps. When users have more than 3000 browses, more than 80% of time gaps between user browsing events is 1s, which suggests abnormal behaviors, i.e., machine generated logs.

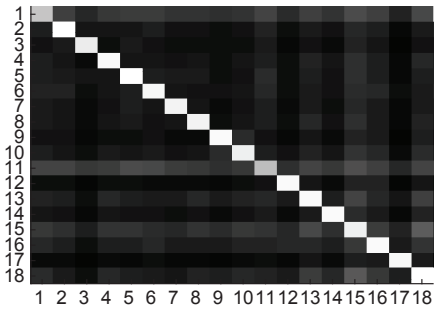

Fig. 6: Average distance of users in one cluster to others, where numbers 1-18 represent each cluster, and lighter color represents smaller distance.

Browsing Entropy: We first examine how diversely that users browse different categories according to (1). Figure 2 shows nearly half of the users (entropy with 0) concentrate on one category, and only 3.6% have a browsing entropy greater than 0.8. This indicates that most of the users focus on a few categories when they are shopping and browsing online.

Repetitive Ratio: Figure 3 shows the browsing repetitive ratio of major categories. We find that there exist significant differences, i.e., *House Appliances* enjoys highest repetitive ratio while *Books & CDs* obtains the lowest repetitive ratio. This suggests that users have different shopping behaviors in different categories, i.e., repetitive or targeted browses.

Co-Clustering of Users and Categories: To investigate prevalent patterns of users' browsing behavior on different categories, we apply the co-clustering algorithm to identify the groups of users and categories simultaneously. We first choose users that have a sufficient and reasonable number of browsing records according to Figure 4 and Figure 5 by focusing on users that browse 100 to 3000 products, finally obtaining 46,366 users.

We obtain 18 clusters (12 major clusters listed in Table 1) and plot the heatmap of average distance among clusters

No.	# Users	Categories	No.	# Users	Categories	No.	# Users	Categories
1	10165	(PC & Office),Ticketing	5	1806	Gifts & Bags	9	1303	Cosmetics
2	9840	Clothing, (Sports & Health), Footwear	6	1470	Maternity & Child	10	359	(E-)Books & CDs
3	8391	Mobile Phones & Accessories	7	1421	Digital Products	11	601	Clocks & Watches
4	7368	Kitchenware, House Decorations, Household Appliances	8	1406	(High-end Brand), Car Accessories	12	455	Toys & Musical Instruments

Table 1: Co-clustering Results.

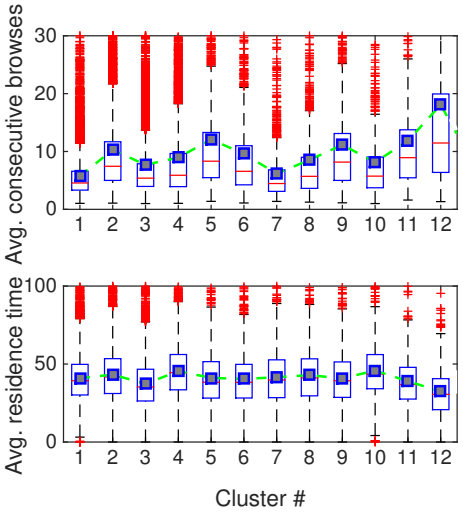

Fig. 7: Average residence time and consecutive browses on the corresponding category in each cluster, where residence time is measured in seconds.

according to (2) in Figure 6, which verifies the effectiveness of co-clustering results. The visualization of the obtained clusters is shown in Figure 8, where we can intuitively observe several enlightening clusters as follow

- **Cluster 1 (Business Usage)**: Users tend to browse office product equipment or ticket booking for business purpose.
- **Cluster 2 (Individual Dressing)**: In this cluster, users always choose the products belonging to *Clothing*, *Sports & Health* and *Footwear* for individual dressing.
- **Cluster 4 (Household Usage)**: Users are browsing products for household usage, including *Kitchenware*, *House Decorations*, *Household Appliances*.
- **Cluster 8 (High-income Group)**: Users in this cluster browsing the products of *Car Accessories* may have individual cars, and are recognized as the high-income group. They at the same time prefer expensive *High-end Brand* products.

Category-based Browsing Behavior Analysis: Based on our analysis about the grouping patterns between users and categories, we study the browsing behaviors of users in each cluster according to Table 1. By averaging user consecutive browsing and residence time in the corresponding category in each cluster, we plot them in Figure 7 with the threshold as 5 minutes to partition sessions. The number of consecutive browses exhibits distinct differences among different clusters. For example, *Toys & Musical Instruments* (Cluster 12) that are popular among children exhibit short residence time but more consecutive browses, which indicates more repetitive browsing; while *(E-)Books & CDs* (Cluster 10) has long residence time, which shows that users are willing to gather more information about the products. In particular, *Ticketing* and *PC & Office* in Cluster 1 attract the least consecutive browses, which suggests that users choose them directly with clear purpose of purchase.

5 Conclusion & Future Work

With a dataset containing users' one-month online shopping and browsing records, we investigate the grouping characteristics between users and product categories, and uncover distinct patterns of browsing behaviors in different categories. Our findings provide valuable insights for e-commerce business to customize its online web shopping system to enhance user experience. As for future work, we would like to study how users preference change over time and how their other online activities are related to their shopping behaviors.

References

1. Tian D, Zhou J, Wang Y, Lu Y (2015) A Dynamic and Self-Adaptive Network Selection Method for Multimode Communications in Heterogeneous Vehicular Telematics.IEEE Transactions on Intelligent Transportation Systems 16(6):3033-3049
2. Tian D, Zhou J, Sheng Z (2017) An Adaptive Fusion Strategy for Distributed Information Estimation Over Cooperative Multi-Agent Networks.IEEE Transactions on Information Theory.pp 99:1-1
3. eMarketer (2016) Worldwide retail ecommerce sales: emarketer's updated estimates and forecast through 2019. http://www.emarketer.com/public_media/docs/eMarketer_eTailWest2016_Worldwide_ECommerce_Report.pdf.
4. Anderson A, Kumar R, Tomkins A, Vassilvitskii S (2014) The dynamics of repeat consumption.International Conference on World Wide Web.pp 419-430
5. Liu CH, Zhang Z, Chen M (2017) Personalized Multimedia Recommendations by Adaptive Feedback Control Frameworks for Cloud-Integrated Cyber Physical Systems.IEEE Systems Journal 11(1):106-117

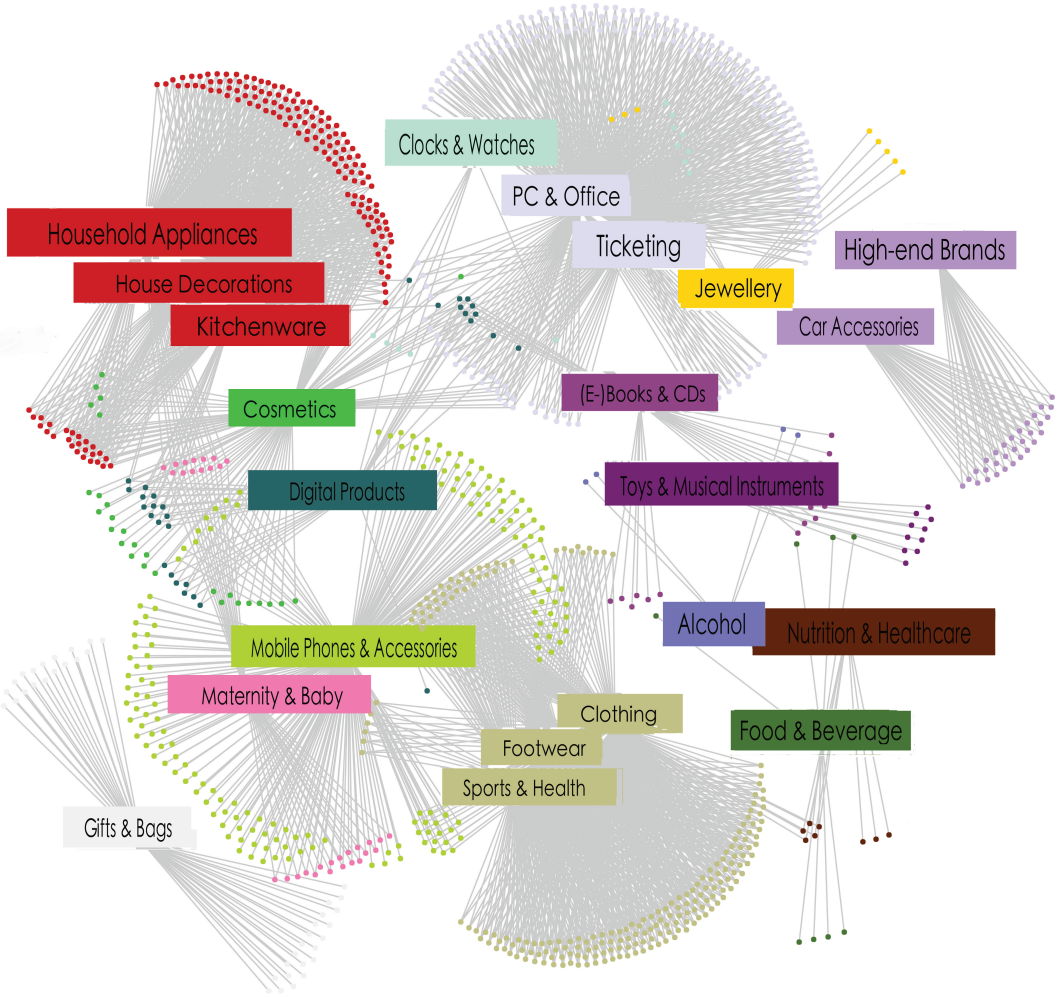

Fig. 8: Diversity of co-clustering results showing user online shopping and browsing behaviors.

6. Zhang Y, Chen M, Huang D, Wu D, Li Y (2016) iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization.Future Generation Computer Systems 66:30-35
7. Zhang Y, Zhang D, Hassan MM, Alamri A, Peng L (2015) CADRE: Cloud-Assisted Drug REcommendation Service for Online Pharmacies.Mobile Networks and Applications 20(3):348-355
8. Benson A R, Kumar R, Tomkins A (2016) Modeling User Consumption Sequences.International Conference on World Wide Web. International World Wide Web Conferences Steering Committee.pp 519-529
9. Keralapura R, Nucci A, Zhang ZL, Gao L (2010) Profiling users in a 3g network using hourglass co-clustering.International Conference on Mobile Computing and Networking, MOBICOM 2010, Chicago, Illinois, Usa, September. DBLP 49:341-352
10. Qiu M, Ming Z, Li J, Gai K, Zong Z (2015) Phase-Change Memory Optimization for Green Cloud with Genetic Algorithm.IEEE Transactions on Computers 64(12):3528-3540
11. Li Y, Chen M (2015) Software-Defined Network Function Virtualization: A Survey.IEEE Access 3:2542-2553
12. Chen M, Hao T, Hwang K, Wang L (2017) Disease Prediction by Machine Learning over Big Healthcare Data.IEEE Access 4:1242-1253
13. Zheng K, Yang Z, Zhang K, Chatzimisios P (2016) Big data-driven optimization for mobile networks toward 5G.IEEE Network 30(1):44-51
14. Qiu M, Sha HM (2009) Cost minimization while satisfying hard/soft timing constraints for heterogeneous embedded systems.Acm Transactions on Design Automation of Electronic Systems 14(2):1-30
15. Li J, Qiu M, Ming Z, Quan G, Qin X, Gu Z (2012) Online optimization for scheduling preemptable tasks on IaaS cloud systems.Journal of Parallel & Distributed Computing 72(5):666-677
16. Chen M, Ma Y, Song J, Lai CF, Hu B (2016) Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring.Mobile Networks and Applications 21(5):825-845
17. Chen M, Yang J, Hao Y, Mao S, Hwang K (2017) A 5G Cognitive System for Healthcare.Big Data and Cognitive Computing 1(1)
18. Zhang Y (2016) GroRec: A Group-centric Intelligent Recommender System Integrating Social, Mobile and Big Data Technologies.IEEE Transactions on Services Computing 9(5):786-795
19. Chen M, Ma Y., Li. Y, Wu D, Zhang Y, Youn (2017) Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System.IEEE Communications 55(1):54-61