Empirical Analysis of Customer Behaviors in Chinese E-Commerce

Jinlong Wang and Ke Gao
School of Computer Engineering, Qingdao Technological University, Qingdao, 266033, China
Email: {wangjinlong, cerro.gao}@gmail.com

Gang Li
School of Information Technology, Deakin University, Victoria, Australia
Email: gang.li@deakin.edu.au

***Abstract*—With the burgeoning e-Business websites, E-Commerce in China has been developing rapidly in recent years. From the analysis of Chinese E-Commerce market, it is possible to discover customer purchasing patterns or behavior characteristics, which are indispensable knowledge for the expansion of Chinese E-Commerce market. This paper presents an empirical analysis on the sale transactions from the 360buy website based on the analysis of time interval distributions in perspectives of customers. Results reveal that in most situations the time intervals approximately obey the power-law distribution over two orders of magnitudes. Additionally, time interval on customer's successive purchase can reflect how loyal a customer is to a specific product category. Moreover, we also find an interesting phenomenon about human behaviors that could be related to psychology of customers. In general, customers' requirements in different product categories are similar. The investigation into individual behaviors may help researchers understand how customers' group behaviors generated.**

***Index Terms*—Customer behaviors, purchasing patterns, power-law distributions.**

I. INTRODUCTION

Going deeply insight into transactions given by e-Commerce Market plays an important role for companies to successfully understand and manage customer relationships as well as flexibly develop marketing strategies [1-3]. In fact, it is particularly interesting and attractive in CRM fields [4-11] and social network analysis [12-13] in recent years. At present a large amount of research papers have been concentrated on famous international e-Commerce websites such as Amazon [14], eBay [15], *etc*. to investigate customer behavior characteristics for helping e-commerce companies improving the service quality. The authors [4] incorporate the RFM (recency, frequency, and monetary) concept [16], to find and generate all RFM sequential patterns from customers' purchasing behavior information to provide references for management decision-making.

However, these analyzed global websites are usually not as popular among Chinese customers. E-Commerce in China has been developing rapidly in recent years, and an active market is growing with a number of burgeoning new Chinese websites, such as 360buy.com, dangdang.com and taobao.com, *etc*. Chinese customers in general prefer to purchase via domestic oriented e-Commerce websites [17], like 360buy etc. Due to the majority of customers in websites like 360buy are Chinese, the related transaction data sets provide a great resource for the discovery of Chinese customer purchasing patterns or customer behavior characteristics. Although some researchers investigate Taobao [18, 19]. This paper presents an empirical analysis on the sale transactions from the 360buy website based on the time interval distributions with human dynamics perspective [20, 21]. For the understanding of customer behaviors we analyzed the time interval of the successive purchase of customer, the time interval of the successive purchase of product and the time interval of customers' corresponding purchase-review time difference. The empirical results could possibly help sellers to get a better understanding of their customers so that better services can be provided.

The rest of the paper is organized as follows: The overview of the dataset is given in Section II. Section III describes the empirical analysis of the dataset. Finally, the conclusion is made in Section IV.

II. OVERVIEW OF THE DATASETS

A customer review is published only after a customer purchased a product. Sequentially, there are time information such as buy time and review time, text information such as pros and cons, and rating information such as individual ratings and total ratings *etc* in it. In order to facilitate the study, we transformed the original format of purchasing transactions which consist of product and review information into two entity-attributes, as Table I shows.

Since our investigation is related to time information which can easily be extracted from reviews, each transaction can be predigested as information including buy time, review time and few other supplementary such as product ID, customer name, and categories. All the data were downloaded from 360buy.com website through crawler programs.

doi:10.4304/jnw.5.10.1177-1184

TABLE I.
ENTITY-ATTRIBUTES OF PURCHASING TRANSACTIONS

Entity	Attributes
Product	Product ID, Product Name, Categories, Brand, Register Time, Total Rating, Market Price, etc.
Review	Customer Name, Buy Time, Review Time, Pros and Cons, Individual Rating, Usefulness, etc.

In order to study the purchasing related time interval distributions, we collected data reference to each product's register time on the website from 2008-12-01 to 2009-12-01. The products are categorized in accordance with the hierarchy on the 360buy website. We have chosen top-layered product categories including Computer Products (CP), Articles of Daily Use (ADU), Mobiles and Digitals (MD) and Domestic Appliance (DA) as datasets for analysis. The product hierarchy on the website is shown in Fig. 1. From the figure we can find that the products are classified gradually from coarse-grained categories to fine-grained categories.

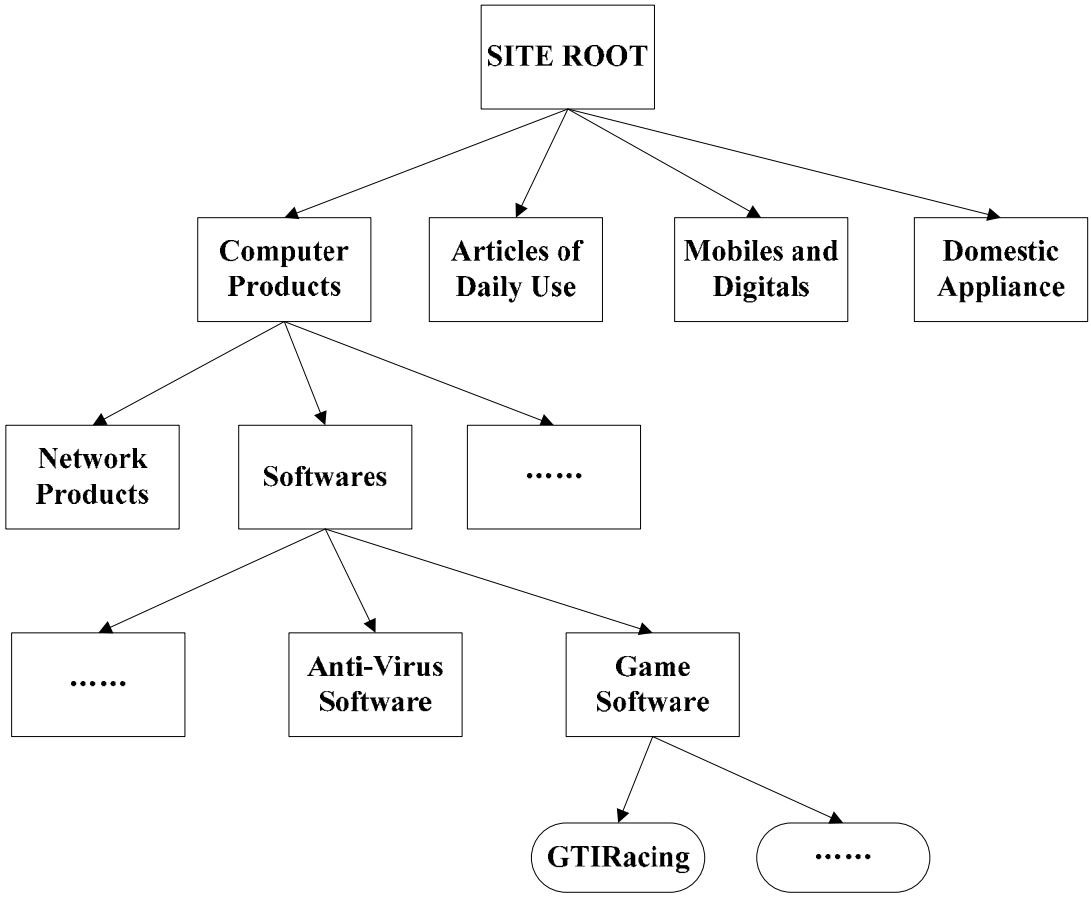

Figure 1. The product hierarchy on 360buy website

The statistics information of these data sets are shown in Table II. The number of active products[1] and the number of active customers[2] are 28620 and 274148, respectively. The purchase time duration is from 2008-12-02 to 2010-01-12. Altogether there were more than 1 million transactions during this period. There are two indices calculated in Table II. AVE_{sc} indicates the average number of purchases of each customer, AVE_{cp} indicates the average number of attracted customers of each product.

TABLE II.
BASIC STATISTICS IN FOUR DATASETS

Datasets	CP	ADU	MD	DA
# Sales	484981	292735	205470	169285
# Products	6904	14054	3497	4165
# Customers	158400	102871	115482	94940
AVE_{sc}	3.06	2.85	1.78	1.78
AVE_{cp}	22.94	7.32	33.02	22.79

The categories of CP, MD and DA are mainly related with 3C electronics and home appliances, which are also getting the best public praise among customers in recent years. The ADU category is later developed on the

[1] The "active" means an event occurs at least once.
[2] Each customer may purchase products in different categories.

website, however the number of products in this category is rather large, and we believe that there are good opportunities in the young category to get great development in the near future. The two indices in Table II can prove the point of view mentioned above: the products in the three categories can successfully attract customers to purchase except ADU. On the other hand, customers are seemed as willing to continuously purchase the products in CP and ADU.

III. Empirical Analysis

For the analysis, we adopt a two-steps strategy as described in [22]: to estimate the parameters of the power-law model toward time intervals, then to calculate the *p*-value for the evaluation of goodness-of-fit between the data and the power law. It is worth to note that, with a narrow time span in datasets a loose *p*-value has to be set to reject implausible hypothesis. Consequently, the hypothesis will be rejected if the *p*-value is less than 0.1.

A. Time Interval Distribution on Customer Successive Purchasing

Time interval on customer successive purchasing (CSP) represents the intensity of customer purchasing behaviors. A customer's impact on others may be related to intensity. CSP distributions are shown in Fig. 2, the power-law fits and the corresponding *p*-value are shown as the 2nd and 3rd columns in Table III. Let x be the frequencies of different time interval occurs. $\Pr(X \geq x)$ represents the probability that a random time interval frequencies is greater than or equal to x. Let α represent a constant parameter of the distribution. We calculated the *p*-value through generating 100 repetitions of synthetic power-law distributions. The plausible power-law distributions are denoted in bold.

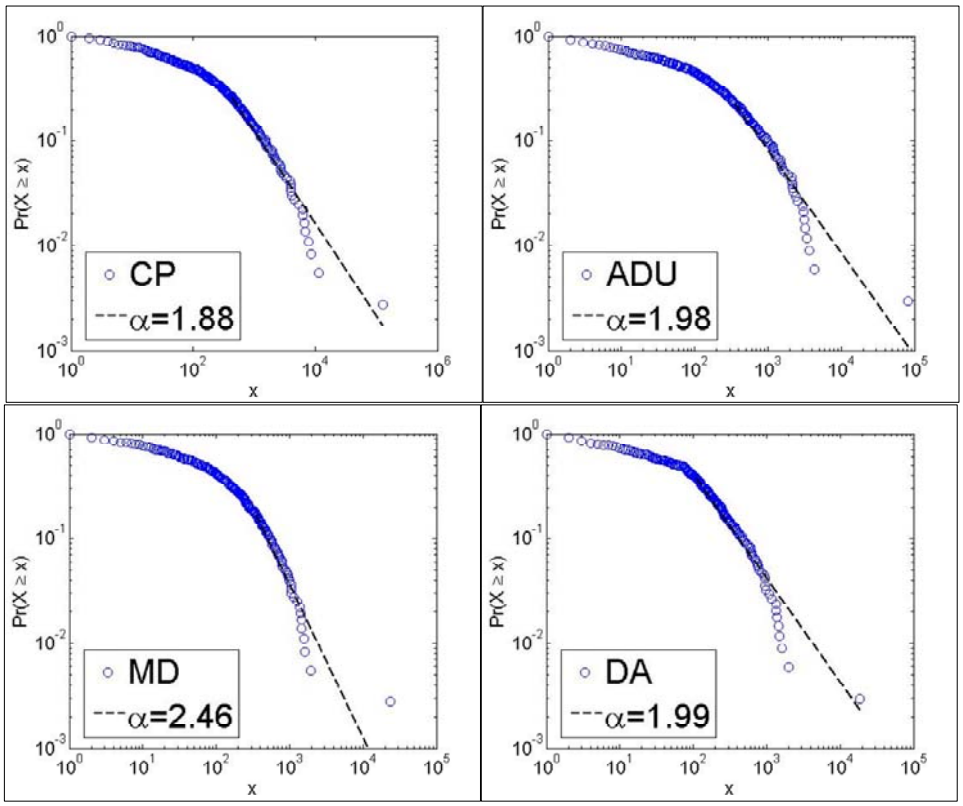

Figure 2. Time interval distributions on customer successive purchase.

TABLE III.
POWER-LAW FITS AND THE CORRESPONDING *P*-VALUE IN VARIOUS TIME INTERVAL DISTRIBUTIONS.

DataSets	CSP		DiffCPR		PSP	
	α	*p*-value	α	*p*-value	α	*p*-value
CP	1.88	**0.21**	1.75	**0.60**	1.50	**0.98**
ADU	1.98	0.03	1.68	**0.17**	1.52	**0.63**
MD	2.46	**0.54**	1.76	**0.14**	1.49	**1.00**
DA	1.99	0.02	1.66	**0.37**	1.50	**0.98**

CSP approximately obeys the power-law distribution within the limit of two to three order of magnitudes both on the Computer Products (CP) dataset and the Mobiles and Digitals (MD) dataset. The statistic $\alpha_{CP}<\alpha_{MD}$ indicates that there are more short time intervals in the transactions related to the Computer Products category.

However it is necessary to note that, the original purchase sequence of a customer includes his purchase transactions from all four product categories, while the analyzed sequence is a subsequence which is related to the interested product category. In this case, comparing with the customers who purchased in the Mobiles and Digitals category, the customers who purchase in the Computer Products category are more likely to stay in original category. Based on this, we believe that the CSP can reflect the loyalty of a customer in a specific product category.

B. Time Interval Distribution on Customer Purchase-Review Time Difference

Time interval on customer purchase-review time difference (DiffCPR) represents how long a customer will review a product after the purchasing. When the time interval is longer, the customer will get more familiar with the product and his/her relevant reviews could be more reliable for others. Statistics are shown in Fig. 3 and Table III, of column 4 and 5. DiffCPR approximately obeys the power-law distribution within the limit of two to three order of magnitudes on all over the four datasets and each α is around 1.7. It indicates that the customers' review patterns are consistent across all product categories. However, there is a discontinued point on each curve in each subfigure of Fig. 3. Through further examination from statistics on embedded points which indicates this part of data, we find that this disconnection phenomenon occurs at adjacent time intervals which are 30 and 31 days respectively. We believe that it is not a noise because fluctuation trends of distributions on the remaining two partitioned data pieces are still consistent with each other. Considering the fact that the numbers of reviews published within a month are far more than that over a month, we believe that this phenomenon could be related to the psychology of customers and this will be one issue for further investigation.

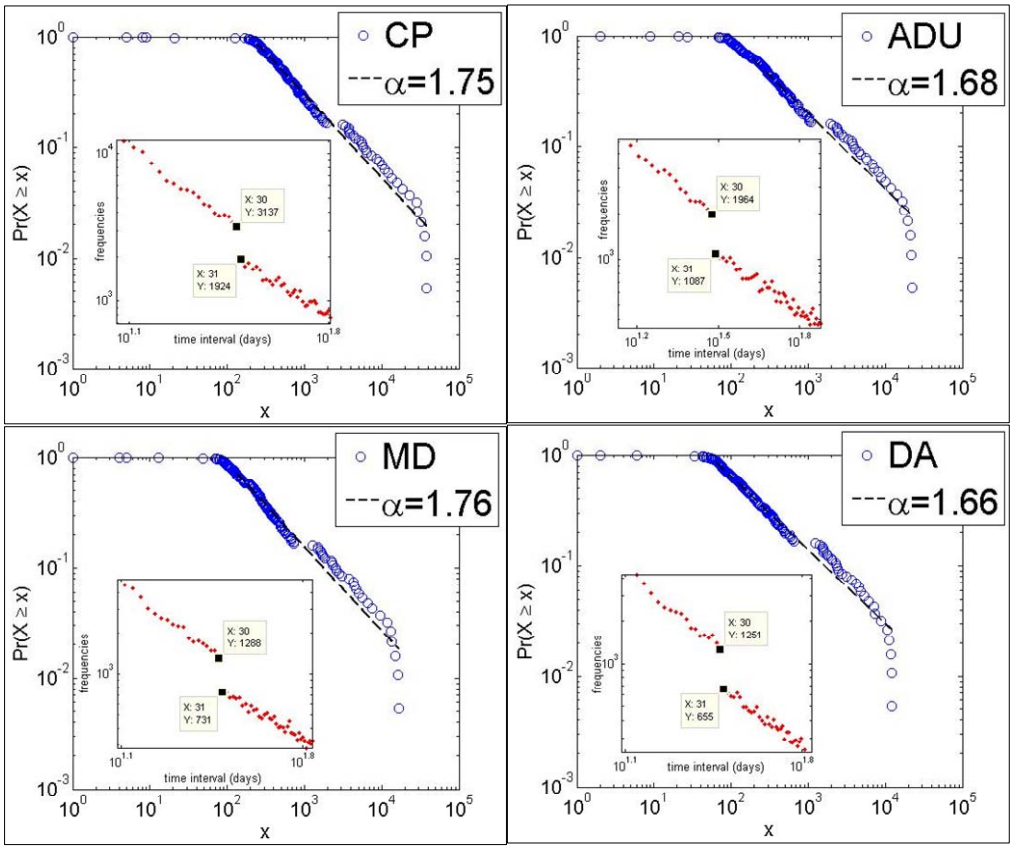

Figure 3. Time interval distributions on customer purchase-review time difference.

C. Time Interval Distribution on Product Successive Purchase

Time interval on product successive purchase (PSP) represents the speed on product sales, it reflects customer's purchasing needs. Statistics are shown in Fig. 4 and Table III, of column 6 and 7. PSP approximately obeys the power-law distribution within the limit of 3 to 4 order of magnitudes on all over the four datasets and each α is around 1.5. Hence, despite the diversity of product attributes in different categories, the distributions of sales are consistent. Consequently, we can conclude from the empirical result, in general, that the purchasing requirements in different product categories between different customers are almost the same.

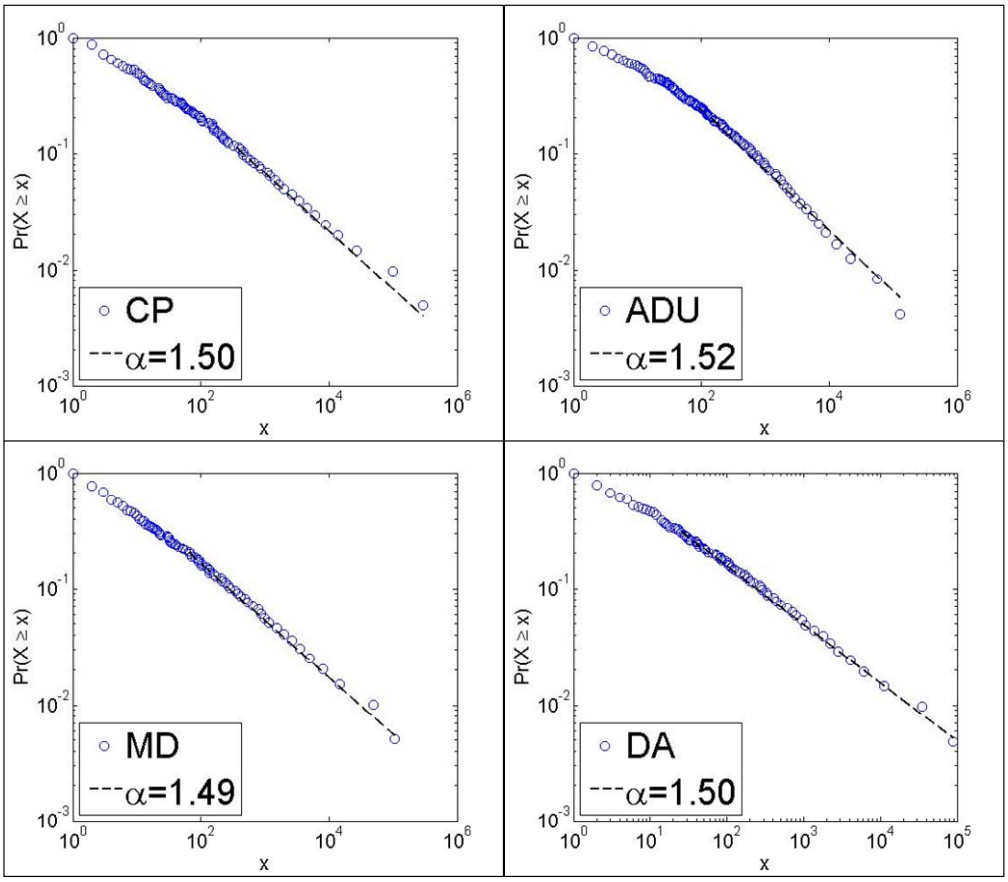

Figure 4. Time interval distributions on product successive purchase.

D. *Individual Product Analysis*

Analyzing some typical products may help us learn whether statistical characteristics are consistent between individuals and product groups. Four top sold products from each of the datasets were chosen in this experiment, and their daily sales statistics are shown in Fig. 5. The ubiquitous "sale peaks" run through the selling period which implies the possible scale-free characteristics. We have planned to empirically study the time interval on PSP and DiffCPR, but PSP are too short to get statistics with time intervals no more than 10^1 order of magnitude.

In this case, we attempt to study DiffCPR, which represents time interval distribution on customer purchase-review time difference. The plausible power-law distributions are denoted in bold. The statistics are shown in Fig. 6 and Table IV. TSP represents top sold products in different product categories. α*@CPC* represents α at corresponding product categories (CPC). DiffCPR approximately obeys the power-law distribution over two orders of magnitudes on Computer Products, Articles of Daily Use, Mobiles and Digitals and each α is around 1.7. Similar to statistics (the data are shown in parenthesis in Table IV) in product category, customers' review patterns are highly consistent in different individuals. Meanwhile, the "disconnection phenomenon" can be reflected though not so obviously. Hence, the statistical characteristics of DiffCPR are consistent between individuals and product categories.

TABLE IV.
POWER-LAW FITS AND THE CORRESPONDING *P*-VALUE IN DIFFCPR,

TSP	α*@CPC*	*p*-value
CP	1.72(1.75)	**0.19**
ADU	1.67(1.68)	**0.14**
MD	1.73(1.76)	**0.16**
DA	1.65(1.66)	0.06

From the above phenomenon we can conclude that the investigation on individual purchasing behaviors may help to us know the possible reason that how customers' group behaviors generated. According to this, since the individual customer's characteristics are more distinctive and easy to obtain, we can choose some of the individuals to study and utilize those characteristics to assist the investigation on customers' group behaviors.

Moreover, combining all the results in the Table III and Table IV, it is not difficult to find that the categories CP and MD are always approximately obey the power-law distribution no matter under study of customer groups or individuals. This means the purchasing dataset of CP and MD are more reliable than others, and they are useful for further studies.

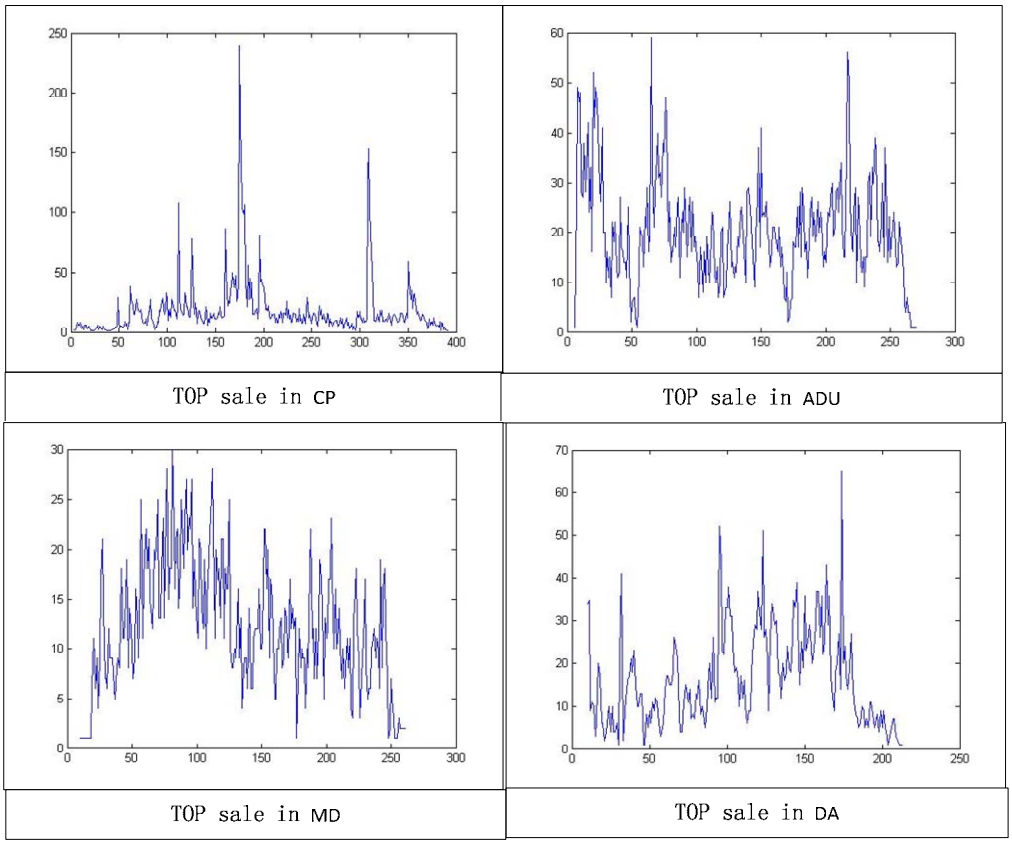

Figure 5. Daily sales statistics of TOP sale products. The X axis denotes customer purchasing time sequence (days) and the Y axis denotes the sales.

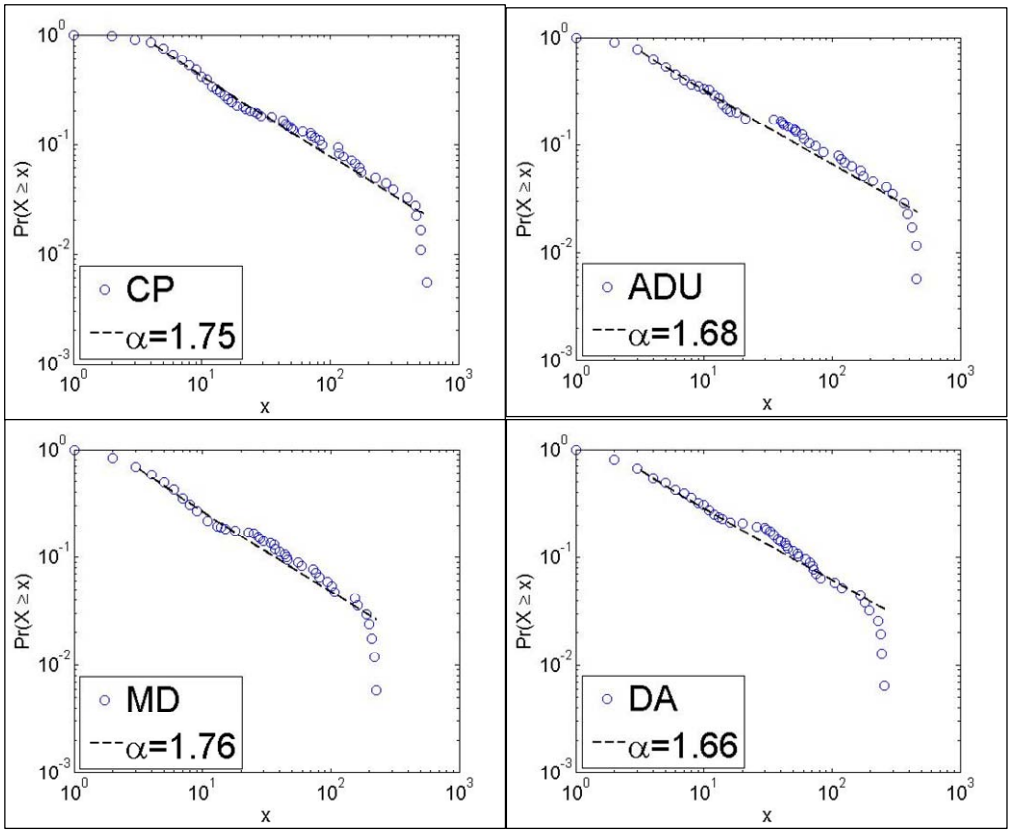

Figure 6. Time interval distributions on customer purchase-review time difference for individuals.

IV. CONCLUSION

This paper presents an empirical analysis on the transaction data of a Chinese e-commerce website (360buy.com) based on the time interval distributions in perspectives of customers. We find that in most situations the time intervals approximately obeys the power-law distribution over two orders of magnitudes. In addition, the time interval on customer successive purchase can reflect how loyal a customer is towards a specific product category. Moreover, an interesting phenomenon about human behaviors is discovered and it may be related to psychology of customers. In general, customers' requirements in different product categories are similar. The investigation into individual behaviors may help researchers to understand how group behaviors generated. As a summary, these findings can help us to further investigate social network analysis and clustering analysis.

For the future work, we will further investigate the user's behavior in e-commerce, compare the difference among different culture customers, and use the obtained behavior characteristics to market segmentations, commodity recommendation, *etc*.

ACKNOWLEDGMENT

This work was partially supported by the National Natural Science Foundation of P.R.China (No.60802066), the Excellent Young Scientist Foundation of Shandong Province of China under Grant (No.2008BS01009) and the Science and Technology Planning Project of Shandong Provincial Education Department (No.J08LJ22).

REFERENCES

[1] Kohavi, R. Mining e-commerce data: the good, the bad, and the ugly. In SIGKDD'01: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 8-13, San Francisco, CA, USA, ACM Press, 2001.

[2] Dibb, S. Market segmentation: strategies for success. Marketing Intelligence and Planning, 1998, 16(7): 394-406.

[3] Turban, E., Lee, J., King, D., and Chung, H. M. Electronic commerce: a managerial perspective. 4th edition. Prentice Hall, Upper Saddle River, NJ, 2006.

[4] Chen, Y. L., Kuo, M. H., Wu, S. Y., and Tang, K.. Discovering recency, frequency, and monetary (rfm) sequential patterns from customers' purchasing data. Electronic Commerce Research and Applications, 2009, 8(5): 241-251.

[5] Cheng, C. H., and Chen, Y. S. Classifying the segmentation of customer value via rfm model and rs theory. Expert Systems with Applications, 2009, 36(3): 4176-4184.

[6] Liao, S. H., Chen, J. L., and Tsu, T. Y. Ontology-based data mining approach implemented for sport marketing. Expert Systems with Applications, 2009, 36(8): 11045-11056.

[7] Huang, S. C., Chang, E. C., and Wu, H. H. A case study of applying data mining techniques in an outfitter's customer value analysis. Expert Systems with Applications, 2009, 36(3): 5909-5915.

[8] Lu, T. C., and Wu, K. Y. A transaction pattern analysis system based on neural network. Expert Systems with Applications, 2009, 36(3): 6091-6099.

[9] Ahn, H. J., Kang, H., and Lee, J. Selecting a small number of products for effective user profiling in collaborative filtering. Expert Systems with Applications, 2010, 37(4): 3055-3062.

[10] Lee, S. L. Commodity recommendations of retail business based on decision tree induction. Expert Systems with Applications, 2010, 37(5): 3685-3694.

[11] Schierz, P. G., Schilike, O., and W. Wirtz, B. Understanding consumer acceptance of mobile payment services: An empirical analysis. Electronic Commerce Research and Applications, 2009, DOI: 10.1016/j.elerap.2009.07.005.

[12] Kiss, C., and Bichler, M. Identification of influencers - measuring influence in customer networks. Decision Support Systems, 2008, 46(1): 233-253.

[13] Li, Y. M., Lai, C. Y., and Chen, C. W. Identifying bloggers with marketing influence in the blogosphere. In ICEC'09: Proceedings of the eleventh International Conference on Electronic Commerce, pp. 335-340, Taipei, Taiwan, ACM Press, 2009.

[14] Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., and Lee, L. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In WWW'09: Proceedings of the eighteenth International Conference on World Wide Web, pp. 141-150, Madrid, Spain, ACM Press, 2009.

[15] Jin, R., Parkes, D. C., and Wolfe, P. J. Analysis of bidding networks in eBay: aggregate preference identification through community detection. In PAIR'07: Proceedings of the AAAI Workshop on Plan, Activity and Intent Recognition, 2007.

[16] Bult, J. R., and Wansbeek, T. Optimal selection for direct mail. Marketing Science, 1995, 14(4): 378-394.

[17] Vuylsteke, A., Wen, Z., Baesens, B., and Poelmans, J. Consumers' online information search: a cross-cultural study between China and Western Europe. http://www.aabri.com/OC09manuscripts/OC09043.pdf. Jan 20, 2010 accessed.

[18] Ye, Q., Li, Y. J., Kiang, M., and Wu, W. F. The impact of seller reputation on the performance of online sales: evidence from taobao buy-it-now (bin) data. SIGMIS Database, 2009, 40(1): 12-19.

[19] Li, D. H., Li, J., and Lin, Z. X. Online consumer-to-consumer market in china - a comparative study of taobao and ebay. Electronic Commerce Research and Applications, 2008, 7(1): 55-67.

[20] Barabási, A. L. The origin of bursts and heavy tails in human dynamics. Nature, 2005, 435(7039): 207-211.

[21] Vazquez, A., Oliveira, J. G., Dezso, Z., Goh, K., Kondor, I., and Barabasi, A. L. Modeling bursts and heavy tails in human dynamics. Physical Review E, 2006, 73(3): 036127.

[22] Clauset, A., Shalizi, C.R., and Newman, M. E. J. Power-law distributions in empirical data. SIAM Review, 2009, 51(4), 661-703.

Jinlong Wang received the Diploma and Ph.D. degree at College of Computer Science from Zhejiang University, China in 2002 and 2007 respectively.

He is currently an Associate Professor at School of Computer Engineering, Qingdao Technological University. His research interests include data mining, machine learning and artificial intelligence.

Ke Gao received his Diploma degree at School of Computer Engineering from Qingdao Technological University, China in 2008.

He is currently a master student at School of Computer Engineering, Qingdao Technological University. His research interests include data mining, text mining and business intelligence.

Gang Li is a lecturer in the school of Information Technology, Deakin University. He completed a PhD in 2005 at Deakin University in the area of data mining, and received the bachelor's degree in computer science from Xi'an Petroleum Institute in 1994, the master by research degree from Shanghai University of Science and technology in 1997.

His research interests include data mining, wireless sensor networks and sentiment mining from multimedia.