International Journal of Business Marketing and Management (IJBMM)
Volume 4 Issue 9 September 2019, P.P. 50-60
ISSN: 2456-4559
www.ijbmm.com

Profile of User Search Behavior and Advertising on Alibaba's Taobao Platform

Min Li[1], Joseph Richards[2]
[1](*Department of Information Systems and Business Analytics, CSU-Sacramento, USA)*
[2](*Department of Marketing, CSU-Sacramento, USA)*

ABSTRACT : *This paper uses the rich features of two data sets obtained from the Alibaba's Taobao online shopping site taobao.com. The page view log data file contains the information about the advertisements displayed after the customer's search using keyword(s). The click log data file contains the information about the advertisements clicked by customers. Both data files contain information specific to customers including their geographic location, the time of search, and the time of clicking an advertisement. Exploratory data analysis reveals the profile of search behavior and advertising strategy employed by Taobao. This analysis will be valuable for replicating the success of Taobao platform in other emerging markets.*
KEYWORDS *- online consumer behavior, online search, sponsored search, taobao*

I. Introduction

Today, China's e-commerce market is the world's most voracious, the Single's Day promotion sales on Alibaba's online platforms exceeded $30 billion in 2018. According to CNBC, Alibaba's Singles Day sales haul easily exceeded the spending by consumers during any single U.S. shopping holiday. By 2020, China's e-commerce market is forecast to be bigger than the size of markets in America, Britain, Japan, Germany, and France combined ("The Alibaba Phenomenon", 2013). Using the Taobao search platform, consumers often obtain information before shopping online, such as product prices, the origin of the goods, product quality, and the seller's credit evaluation. The search engine also records the query process and any clicks by the customer on advertisements or search links displayed (Wei et al., 2014).

The plethora of innovations that are emerging from China calls for a closer look to study their specific features, and perhaps by doing so will offer valuable lessons for other emerging markets. This paper uses the customer search data from Taobao.com, a consumer-to-consumer portal and one of the principal arms of Alibaba, for analyzing the search advertising business carried out by Taobao. Customer search behavior in conjunction with the search advertising carried out by the portal in response to search and browsing, provide a rich source of data for analysis and important insights.

II. The Taobao platform

The users search for products on Taobao (taobao.com) by entering keywords. After a user searches for a product by entering a keyword, four columns of pictures and texts relevant to this product are displayed in the central area of the webpage. Two additional columns, on the right and at the bottom, display product advertisements from various merchants who have bid this keyword entered by the customer. Discounted prices for the advertised products are usually shown in these columns with the original list (usually higher) prices crossed out. Taobao's search engine matches the keywords entered by customers to the keywords (*bidwords* in Taobao's terminology) merchants have bid for online advertisements. The search engine must determine which advertisements to display on the right side and at the bottom of the page, and the order and position of the advertisements. The advertisement receiving the highest bid from the merchants is usually placed at the top of the list. Merchants pay Taobao the bid price each time a potential customer clicks the advertisement. Thus, the advertising mechanism on Taobao.com is similar to Google's AdWords. In addition to sponsored search advertising, Taobao also provides other services including display advertising (contextual advertising) and

behavioral targeting. Display advertising involves showing advertisements such as banners on a website aiming to drive the traffic to advertisers' own site. Behavioral targeting predicts customers' interests based on their past online behavior and other features (see Svensen, et al. 2011). Then customers are divided into segments of similar interest for targeted advertising campaigns. All these services allow merchants and advertisers to reach a large number of target customers with the highest purchase intent for the advertised products.

This paper is exploratory in nature and complements the work done by the authors in a related paper (see Richards and Li, 2018) where the authors' focus was to infer strategic insights that could be tested by the data. This paper presents exploratory analysis of the data and suggests potential managerial implications which are of immediate concern for e-commerce-oriented internet platforms.

The next section describes the data and its analysis. In the data analysis part, the profile of user search and specific features of Taobao's search advertising strategies are presented. The paper concludes with a discussion on the managerial implications and avenues for further exploration.

III. Data and Analysis

In sponsored search advertising, a customer searches a term using a search engine and the search results (links) including those sponsored by advertisers are returned. Customers' clicking on links sponsored by advertisers results in payments to search engines. Thus, it is in the interest of search engines to maximize the click-through rate (CTR), defined as the ratio of the number of clicks an advertisement received and the number of times the advertisement was shown. Advertisement features include bid phrases (bid words, keywords), the title of the advertisement, the text of the advertisement, the URL of the landing page, the landing page, and the advertiser information such as account, campaign, advertisement group, and the advertisement. Query features include the search keywords and related expansion. Context features include display and geographic locations, time, search history, and other customer data. Many of these data features appear in the two data sets from Taobao described below.

The data from Taobao's paid search ranking (pay-for-performance) system are used for this study. Two data sets were sampled at one author's request by data scientists at Taobao between 2 pm and 2:05 pm on May 16, 2013, recording online customers' behavior during this time period. The short 5-minute sampling period was chosen so that the data file size is manageable. These data files include Chinese characters taking up several gigabytes of storage space. The variables captured in the two data files, the page view log file and the click log file, are described in detail below.

Page View Log file

Table 1 describes the important variables in this data file. This file recorded the information related to the advertisements displayed to the customer after the customer's search using keyword(s) from 2 pm to 2:05 pm on May 15, 2013. The file contains 997,407 observations and 65 variables. About one third of the variables are ID-type variables referring to cities, provinces, keywords merchants placed bids on, customer queries, advertising customers, advertisement groups, etc. Variables such as *adhighestprice*, *adprice*, *adrelativeposition*, *adscoretag*, and *adsnumperpage* are explained in **Table 1**, and are similar to the online advertisement features described in Graepel et al. (2010). Three variables, *pvtime*, *city*, and *province* show the time and location of the customer when viewing this page. These belong to the context features also described in Graepel et al. (2010). Some of these variables are used to track advertising customers, specific advertisements, or traffic related to the websites. The variable *rawquery_keyword* contains the search keywords and its related expansion. The variable *bidword* contains words that are a part of the variable *rawquery_keyword*. Both variables belong to the query features described in Graepel et al. (2010). Additional explanation of select variables and the information they capture are given in **Table 1**.

Frequencies of the variable *adsnumperpage* are reported in **Table 2**, ordered from the most frequent to the least frequent, after sorting the data by customers (*acookie*) and the keywords (*rawquery_keyword*) these customers searched, removing duplicate values. Sorting by these two variables and removing the duplicate values were necessary so that each search result webpage corresponds to only one value for the variable

adsnumperpage. The most common number of advertisements per page is 13 (37%) and the next one is 3 (18.3%). The range observed is 59. The number of advertisements commonly displayed shows a wide range from 3 to 13. This certainly is on account of the user, the type of product searched for that session, and the length of the search session. It also shows a high degree of customization of advertisements to the user. The number of advertisements shown in response to user search will likely depend, among other factors, on the type of product searched, whether the user uses a query word, and the time spent in a search session. Exploration of these relationships is a topic for future research. Knowledge of these relationships will reveal the specific strategies employed by Taobao on the platform.

The variable *adprice* contains the prices of products displayed in the sponsored advertisements on the right and at the bottom of the page, for each customer identified by the variable *acookie*. The variable *adhighestprice* captures the highest list price or regular price of the advertised product from which the price is marked down to *adprice*. For most advertisements, *adhighestprice* is crossed out and *adprice*, i.e., the discounted price, is displayed next to it. The *adhighestprice* ranges from 5 to 9,900 Chinese Yuan, with 164 as the average and 122 Yuan as the median price. The histograms of both variables, *adprice* and *adhighesprice*, are shown in Figures 1 and 2. A few observations may be made about these histograms. The long-tailed distribution observed in these price histograms demonstrates that the advertisements shown are for products from the low-end to the high-end of the price spectrum. Avenues of further inquiry include whether the price dispersion for advertised products within a category differs between product categories, whether the dispersion depends on the user using a query key word for search, and whether the dispersion and the length of search session are related.

Fig. 3 displays the histogram of *adscoretag*, recording the number of clicks on the advertisement. The range of this variable value is from 0 to 999, with a mean of 373 and median of 311. The first quartile is 207 and the third quartile is 481. The histogram in Figure 3 shows that most advertisements resulted in fewer than 500 clicks. The histogram shows a long-tailed distribution proving that some types of advertisements stand out. An avenue for further research is to study the characteristics of such advertisements and figure out the what factors might have caused a spike in the number of clicks.

Fig. 4 displays the histogram of the percentage by which *adprice* is marked down from *adhighestprice*, i.e., the discount rate based on *adprice* and *adhighestprice*. There is no discount for about half of the advertisements, and 25% of the advertisements have a markdown of at least 5%. An important insight would be to know whether the amount of price discounts had any effect on the click through rates (CTR) and whether the amount of discount varied across product categories.

Tables 3 and **4** list the top 10 cities and provinces where customers are located. Almost half of the customers are in the coastal provinces (Guangdong, Zhejiang, Jiangsu, Shandong, Beijing, Shanghai), cities that are relatively more developed compared to the rest of China. Most of the top 10 cities in **Table 4** are also along the coast. Some customers are from foreign countries. Further analysis could explore if user search behavior and Taobao's advertising strategy show marked difference across geographical regions. If user behavior is dependent on geographical status then that information could be made part of the overall customization of advertising strategy.

Click Log File

The second data file contains click logs between 2 pm and 11:41 pm on May 15, 2013. There are 312,660 observations and 90 variables. In **Table 5** the variables relevant to this study are described.

The variable *fromdomainname* records the origin or URL from which the advertisement originated. The frequency of this variable is shown in **Table 6**. Over 43% of the clicks originated from Taobao's main search engine s.taobao.com. The next frequent domain is etao.com, a vertical shopping search engine in China once powered by Microsoft's Bing search engine. The third domain is list.taobao.com, which is market list by category of products. The rest of the domains are other specialized shopping websites run by Taobao. Clearly,

Taobao's main search engine brings the most number of clicks (43.81% of all clicks). The observed pattern of advertisement origin shows that most of the advertisements originate from Taobao owned or controlled affiliates. There is apparently a high degree of cross selling through advertisements across the platforms as a result. The strategy, it seems, is to use the various websites under the Taobao control umbrella to channel user clicks to the main e-commerce platform of Taobao.

Frequencies of the number of advertisements per page (*adsperpage*) are listed in **Table 7** ordered from the most frequent to the least frequent. Close to half (44.31%) of the clicked advertisements are part of 13-advertisement combination per page (8 advertisements on the right and 5 advertisements at the bottom). Over 12% of the clicked advertisements are part of 24-advertisements combination per page. Other values of *adsperpage* correspond to far fewer number of clicks. It is interesting to note that almost half (44.31%) of the users who clicked an advertisement were shown the standard 13 advertisement combination per page. It appears that the platform is not making a more customized approach to user characteristics that would in turn result in a wider range of the number of advertisements than that is observed.

Tables 8 and **9** list the top cities and provinces the customers who clicked the advertisements are located. As is in the page view log file, most customers are located in coastal provinces and cities. As noted earlier, further analysis could explore if user search behavior and Taobao's advertising strategy show marked difference across geographical regions. If user behavior is dependent on geographical status then that information could be made part of the overall customization of advertising strategy.

Table 10 lists the frequencies of the relative position (the variable *relativeposition*) of the clicked advertisements on a page. This variable assumes integer values between 1 and 75, values corresponding to relative positions of the advertisement compared to other advertisements on the webpage, as designated by Taobao on its search result webpage. Based on the frequencies in **Table 10**, it is clear that the relative position of 1 (the top position on the right) corresponds to the most prominent or desirable position on the webpage as it corresponds to the greatest number of clicks. There is clear correlation between the position of advertisement and the frequency of clicks. As the relative position increases, the corresponding number of clicks decreases, clearly a result of less prominent positions on the web page. This result demonstrates that specific positions of advertisements on the webpage are instrumental in influencing the number of clicks, a matter of great interest for advertisers who want to maximize click-through revenues.

IV. Conclusion

In conclusion, the data analysis presented in this paper is one of the first in its kind to reveal the search based adverting strategy of Taobao, based on actual transactional data from Taobao. China has emerged as the world's largest e-commerce market, and the analytical sophistication employed in monetizing customer search in the e-commerce realm will only increase and this paper contributes in some measure to our understanding of this important market. We have discussed the potential avenues to explore additional insights based on the analysis existing data as well as with future data from Taobao. This paper presents a rather preliminary analysis of field data from the Taobao platform, which by itself is an important contribution because of the paucity of transactional data from actual e-commerce platforms made accessible to researchers and to the public. The purpose is to highlight the potential research avenues that could be explored, which we feel would be immensely relevant to emerging market businesses engaged in some form or other in following the Taobao's business model.

V. Figures and Tables

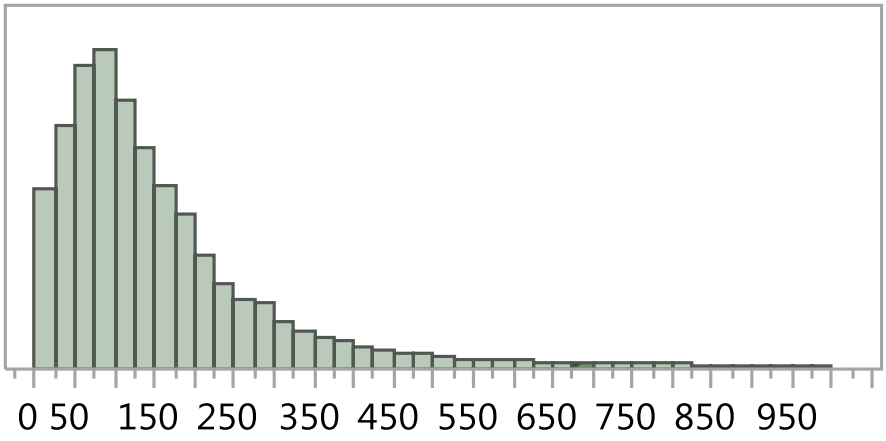

Figure 1. Histogram of the variable *adprice*, the price of the merchandise displayed in the advertisements.

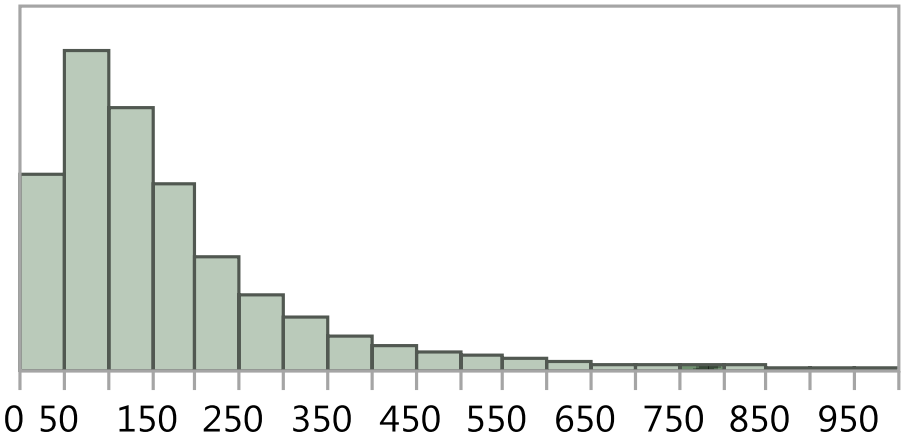

Figure 2. Histogram of the variable *adhighestprice*

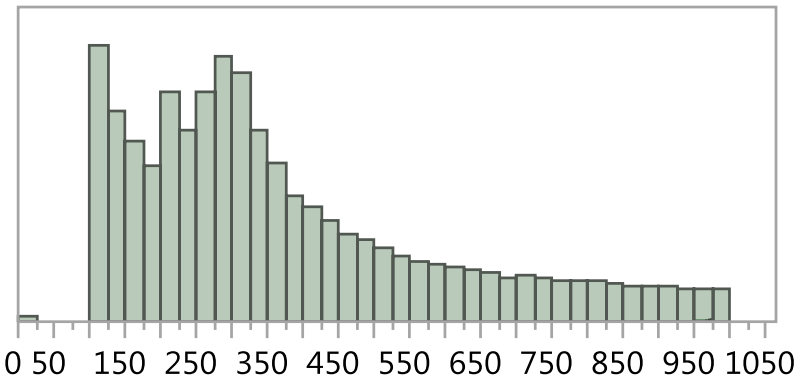

Figure 3. Histogram of the variable *adscoretag*, the number of clicks received by the advertisements.

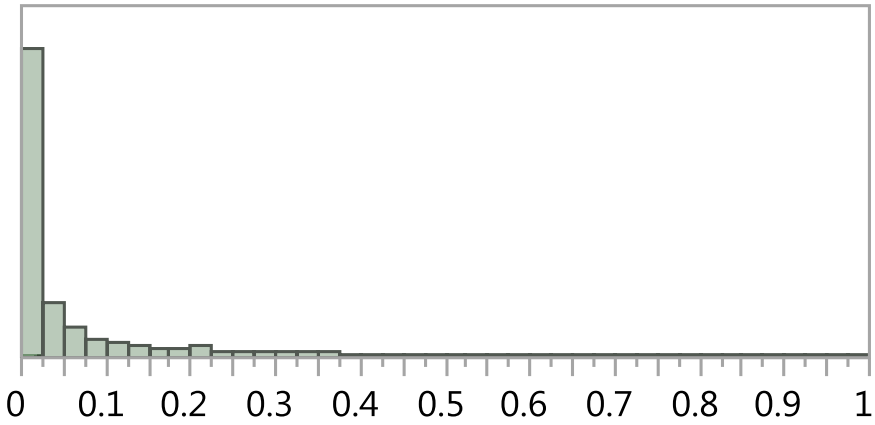

Figure 4. Histogram of the percentage by which *adprice* is marked down from *adhighestprice*.

Table 1. Description of variables in the Page View Log file

Variable Name	Description of Variable
acookie	A variable identifying the customer on the website.
adhighestprice	The highest price of the advertised product from which the price is marked down.
adprice	The price of the advertised product (the marked down price from *adhighestprice*).
adrelativepostion	The position of the ad on the page ranging from 1 to 75 with 1 indicating the best position.
adscoretag	The number of times an ad has been clicked.
adsnumperpage	The number of advertisements per page.
bidword	The keyword a merchant placed the highest bid on.
city	The city in which the customer is located.
province	The province in which the customer is located.
pvtime	Time when page view took place; used to calculate time until click.
rawquery_keyword	The keyword a customer searches for on Taobao.com. The words for this variable contain the bidword. A search for these keywords activates advertisements for display. If the customer clicks on any of these advertisements, Taobao gets paid.
sessionid	The ID of each session of page view. Sponsored advertisements displayed on the right after a search are considered as part of one distinct session ID while sponsored advertisements displayed at the bottom from this search are considered part of another distinct session ID. It can be used to link to the ID in the click log files (see Section 2.2) variable, *pvid*. This variable was generated randomly with 32 alphanumeric characters in length.

Table 2. Frequency of the number of advertisements per page (adsnumperpage), ordered from the most frequent to the least frequent.

adsnumperpage	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
13	13827	37.06	13827	37.06
3	6834	18.32	20661	55.38
16	4099	10.99	24760	66.36
5	3130	8.39	27890	74.75
6	1902	5.10	29792	79.85
24	1771	4.75	31563	84.60
32	1320	3.54	32883	88.13
20	1045	2.80	33928	90.94
27	727	1.95	34655	92.88
1	685	1.84	35340	94.72
12	645	1.73	35985	96.45
7	451	1.21	36436	97.66
4	179	0.48	36615	98.14
15	148	0.40	36763	98.53
8	143	0.38	36906	98.92
40	137	0.37	37043	99.28
10	91	0.24	37134	99.53
30	78	0.21	37212	99.74
2	35	0.09	37247	99.83
50	23	0.06	37270	99.89
60	18	0.05	37288	99.94
9	11	0.03	37299	99.97
11	11	0.03	37310	100.00
13	13827	37.06	13827	37.06

Table 3. Frequencies of the province origin of the customers, top 10 provinces.

province	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Guangdong	3042	11.3	3042	11.3
Zhejiang	2604	9.67	5646	20.97
Jiangshu	2424	9	8070	29.98
Shandong	1697	6.3	9767	36.28
Beijing	1665	6.19	11432	42.47
Shanghai	1313	4.88	12745	47.35
Sicuan	1049	3.9	13794	51.24
Henan	990	3.68	14784	54.92
Hebei	989	3.67	15773	58.59
Fujian	968	3.6	16741	62.19

Table 4. Frequencies of the city origin of the customers, top 10 cities.

city	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Beijing	1665	6.19	1665	6.19
Shanghai	1313	4.88	2978	11.06
Guangzhou	798	2.96	3776	14.03
Hangzhou	751	2.79	4527	16.82
Shenzhen	715	2.66	5242	19.47
Foreign Countries	624	2.32	5866	21.79
Chengdu	549	2.04	6415	23.83
Shuzhou	548	2.04	6963	25.87
Chongqing	500	1.86	7463	27.72
Tianjin	487	1.81	7950	29.53

Table 5. Description of variables in the click log file and a sample of fifteen observations (the values under city and keyword are English translation of Chinese characters).

Variable Name	**Description of Variable**
pvid	An identification (ID) variable similar to *sessionid* in the page view log file. It can be used to link to the ID in the click file's variable, *sessionid* (see Table 1). This variable was generated randomly with 32 alphanumeric characters in length. An exact match indicates the customer clicked the ad after viewing the ad on the page.
clicktime	Time when ad was clicked. It can be used to calculate time until click.
clickprice	The price of the clicked product.
clickpriceorigin	The original price of the clicked product.
highestprice	The highest price of the advertised product from which the price is marked down.
fromdomainname	Indicate the part of taobao.com the click originated from. There are a number of sites of taobao.com, e.g., taobao's search engine or tmall (online store fronts at taobao.com).
clickcity	The city in which the clicking customer is located.
clickprovince	The province in which the clicking customer is located.
keyword	The keywords a customer searches for on the site.
customerid	ID assigned to merchandise advertisers
clickcookie	A variable identifying customers who clicked on the advertisements.
pvtime	Time when page view took place; used to calculate time until click.

Table 6. Frequencies of the origin or URL (fromdomainname) of the advertisements, domains with at least one percent of the total advertisements.

fromdomainname	Frequency	Percent	Cumulative Frequency	Cumulative Percent
s.taobao.com	133306	43.81	133306	43.81
tao.etao.com	51171	16.82	184477	60.63
list.taobao.com	18020	5.92	202497	66.56
re.taobao.com	17212	5.66	219709	72.21
trade.taobao.com	14435	4.74	234144	76.96
s8.taobao.com	14160	4.65	248304	81.61
www.taobao.com	6149	2.02	254453	83.63
list.tmall.com	5870	1.93	260323	85.56
s8.m.taobao.com	4977	1.64	265300	87.2
search.taobao.com	4838	1.59	270138	88.79
favorite.taobao.com	4558	1.5	274696	90.29
rate.taobao.com	3874	1.27	278570	91.56
nvzhuang.taobao.com	3814	1.25	282384	92.81
s.m.taobao.com	3814	1.25	286198	94.07

Table 7. Frequency of the number of advertisements per page (adsperpage) in the click log file, ordered from the most frequent to the least frequent.

adsperpage	Frequency	Percent	Cumulative Frequency	Cumulative Percent
13	138535	44.31	138535	44.31
24	38406	12.28	176941	56.59
3	21426	6.85	198367	63.44
16	18254	5.84	216621	69.28
27	13457	4.3	230078	73.59
5	10440	3.34	240518	76.93
40	9230	2.95	249748	79.88
20	7352	2.35	257100	82.23
12	7249	2.32	264349	84.55
7	7195	2.3	271544	86.85
6	6996	2.24	278540	89.09
32	6575	2.1	285115	91.19
10	5316	1.7	290431	92.89
1	5250	1.68	295681	94.57
30	3798	1.21	299479	95.78
4	2376	0.76	301855	96.54
15	2260	0.72	304115	97.27
9	2129	0.68	306244	97.95
18	1837	0.59	308081	98.54
8	1745	0.56	309826	99.09
11	1303	0.42	311129	99.51
\N	433	0.14	311562	99.65
60	365	0.12	311927	99.77
14	354	0.11	312281	99.88
2	243	0.08	312524	99.96
50	75	0.02	312599	99.98
55	51	0.02	312650	100
28	6	0	312656	100
100	3	0	312659	100
23	1	0	312660	100

Table 8. Frequencies of the province origin of the customers, top 13 provinces.

clickprovince	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Guangdong	32052	10.25	32052	10.25
Jiangshu	30204	9.66	62256	19.91
Zhejiang	30149	9.64	92405	29.55
Shandong	21800	6.97	114205	36.53
Beijing	16279	5.21	130484	41.73
Shanghai	14086	4.51	144570	46.24
Henan	13122	4.2	157692	50.44
Sicuan	12510	4	170202	54.44
Hebei	12456	3.98	182658	58.42
Liaoning	11659	3.73	194317	62.15
Hubei	11148	3.57	205465	65.72
Fujian	10899	3.49	216364	69.2
Anhui	9913	3.17	226277	72.37

Table 9. Frequencies of the city origin of the customers, top 10 cities.

clickcity	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Beijing	16279	5.21	16279	5.21
Shanghai	14086	4.51	30365	9.71
Hangzhou	8359	2.67	38724	12.39
Shenzhen	7496	2.4	46220	14.78
Guangzhou	7159	2.29	53379	17.07
Shuzhou	6445	2.06	59824	19.13
Tianjin	6227	1.99	66051	21.13
Chengdu	5852	1.87	71903	23
Chongqing	5657	1.81	77560	24.81
Nanjing	4927	1.58	82487	26.38

Table 10. Frequencies of the relative positions (*relativeposition*) of the clicked advertisements on a page, top 16 positions.

relativeposition	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	42904	13.72	42904	13.72
2	38422	12.29	81326	26.01
3	34769	11.12	116095	37.13
4	22854	7.31	138949	44.44
5	19331	6.18	158280	50.62
6	17033	5.45	175313	56.07
7	15403	4.93	190716	61
10	15187	4.86	205903	65.86
11	15048	4.81	220951	70.67
9	14413	4.61	235364	75.28
8	13715	4.39	249079	79.66
12	13294	4.25	262373	83.92
13	11620	3.72	273993	87.63
14	4235	1.35	278228	88.99
15	3911	1.25	282139	90.24
16	3663	1.17	285802	91.41

References

[1] Graepel, T., Candela, J., Borchert, T., Herbrich, R., Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In: Proceedings of the 27th International Conference on Machine Learning, 2010.

[2] Svensen, M., Xu, Q., Stern, D., Hanks, S., Bishop, C., "Broad vs Narrow: Modelling Strategies for Online Behavioural Targeting," in Proceedings of the Fifth International Workshop on Data Mining and Audience Intelligence for Advertising (ADKDD), San Diego, USA, ACM Press, 2011.

[3] Richards, J., and Li, Min., The Chinese E-commerce Search Advertising Business: A Case Study of Taobao. Contemporary Management research, 14(2), 2018, 121-142.

[4] The Alibaba Phenomenon; E-commerce in china. (2013, Mar 23). The Economist, 406, 15.

[5] Wei, D., Geng, P., Ying, L., & Shuaipeng, L. (2014, May 31 2014-June 2 2014). A prediction study on e-commerce sales based on structure time series model and web search data. Paper presented at the 26th Chinese Control and Decision Conference (2014 CCDC).