International Journal of Scientific Research in Engineering (IJSREM)
Volume: 04 Issue: 05 | May -2020 ISSN: 2582-3930

Analysis of Top Users in Video Applications By Big Data

C Sarishma [1], P Balaji[2]

[1]PG Scholar, Dept of MCA, SIETK PUTTUR.

[2]Asst. Professor, Dept of MCA, SIETK, PUTTUR, AP.INDIA 517583.

***Abstract*—From an Internet service provider's prospective, the increasing popularity of mobile devices and broadband Internet has created new business challenges: more diverse competitions and more volatile customer behaviors. Therefore, to accurately respond to the changing customer demands, using big data to analyze existing and potential customers has become a trend among businesses in designing marketing plans and products. In this research, heavy users of 12 top video apps are identified using the network connection records in Chunghwa Telecom, Taiwan. Together with hundreds of previously extracted user features, chi-squared and ANOVA tests are performed to find features that have statistically significant differences between heavy and non-heavy users. Such profiling results can be used to design corresponding marketing plans.**

Keywords-user profiling; ţig data

I. INTRODUCTION

The increasing popularity of mobile devices and broadband Internet has produced a more massive amount of data than ever, and this popularity has also changed consumers' lifestyle to rely more on the Internet for daily activities. From a service provider's perspective, the diversity of potential competitions and the volatility of customer behaviors have become the new business challenges. Therefore, a better strategy is required to accurately respond to the changing customer demands.

Nowadays, the marketing strategy has shifted from mass marketing and differentiated marketing to "target marketing" in order to attract the attention of a specific group of customers. A company needs to collect and analyze customer data to identify the characteristics of its target customers. Given the huge amount data produced from the Internet, using big data to analyze existing and potential customers has become a trend among businesses in designing marketing plans and products.

The purpose of this research is to find potential customers of Chunghwa Telecom's online video service by studying the user behaviors in several competing video apps. The results are used by the marketing division to design promotion plans.

The scope of this research is the mobile and broadband customers in Chunghwa Telecom, Taiwan. By using the network connection records in Chunghwa Telecom, together with hundreds of previously extracted user features in the data warehouse, the proposed Chunghwa Telecom big data analysis platform [1] [2] has identified statistically significant features between heavy and non-heavy users of top video apps. Such profiling results can be used to design corresponding marketing plans to boost sales [3].

II. BACKGROUND

For a long time, the Internet has become a major video platform for consumers worldwide. Therefore, online video services have become a new market that telecommunications companies actively develop. In 2014, Chunghwa Telecom launched CHT Video, a public service offering a variety of online video content for subscribers, including live videos, movies, dramas, sports, etc. As a paid service for most of the video content, keeping existing customers and finding potential customers are two very important issues.

Since CHT Video is a relatively new service, the usage data are still limited and unstable. We hope to study the user behaviors of competing video apps to assist the promotion of our own service.

III. RELATED WORK

Mass marketing and differentiated marketing used in the past are focused on non-specific customers. Although they contact a wide range of customers, low proportion of them are interested in these services and benefits are not significant. Therefore, some experts are going to develop the method of precision marketing.

Since the well development of data mining techniques, companies have been able to find out information hidden within the consumption records, personal data and other information. Based on these customer characteristics and preferences, service providers can offer the right products to the right customers now. For example, the decision tree analysis can be used to predict whether the customers will use mobile internet access and the "mPro" service offered by Chunghwa Telecom [4]. In another research [5], the decision tree analysis is used to identify potential customers for wireless broadband services. According to these studies, someone of these variables have much more importance within all of the related dependent variables.

Since the decision tree algorithm has its limitations, for example the accuracy might be reduced when too many categories exist [6]. Therefore, a modified heuristic method

to construct the binary decision tree for improving the accuracy of classification had been proposed [7].

In order to avoid the problem caused by using a single machine learning algorithm, the Chunghwa Telecom big data analysis platform is capable of performing neural network, decision tree and logistic regression for data classification analysis. A lift chart is presented in the platform to determine which algorithm has the best hit rate.

IV. THE ANALYSIS FRAMEWORK

The basic idea of this research is to provide valuable information services through data mining technique. The Chunghwa Telecom big data analysis platform has hundreds of data nodes which can provide big data computing technology, such as MapReduce, HDFS and Hive, for the collected customer data, account information, communication records, network equipment records. The platform captures and selects the information depending on the type, and then stores them in the Data Warehouse. Currently, the amount of data compressed is about 3.4TB average daily. The platform is integrated with existing IT infrastructure to provide professional knowledge analysis tools, as shown in Fig. 1.

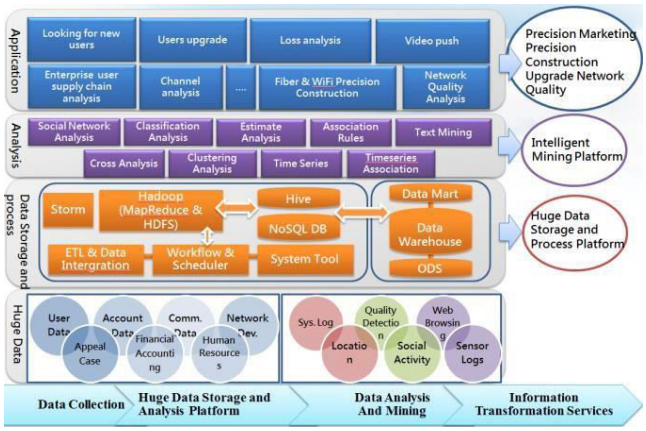

Figure 1. The Analysis Framework.

In order to extract useful information, data should be analyzed by many different dimensions based on the specific requirement. This research, built on this architecture, uses chi-square analysis (Chi-square Test) of cross-analysis to obtain correlation relationship between the customers and service attributes, for example, the relationship within the "heavy users" (defined in Section V.A) of popular video service and video services provided by Chunghwa Telecom.

Through the ANOVA analysis, the relationship between various business attributes and specific customers could be found out. For example, the analysis could show whether there are obvious differences between the access rate which heavy users applied and the actual amount of data transferred.

Then, we can apply logistic regression algorithm to get the relationships between heavy users and business attributes, for example, to find out interest classification rules for heavy users by classification analysis.

Finally, we provide the analyzed results to the business owner to create the customer profile and make suitable marketing plan.

V. EXPERIMENTS AND ANALYSIS

In this section, we describe the data source of our experiments, the definitions of top video apps and heavy users, and present the profiling results obtained from the analysis framework introduced in Section IV.

In this research, the "top" video apps are defined based on download counts, rating scores, and number of ratings from Google Play and/or App Store. We define 12 video apps that possess high levels of such characteristics as the "top" video apps.

In the data collection process, we use the network devices in Chunghwa Telecom to collect user connection records. In this research, we observe 8 weeks of data, from January to February 2015. By using big data computations, we can obtain only those records related to the top video apps we defined. On average, the size of raw data we process per log date is 560 GB.

Next, we label the "heavy users" as positive samples in the data warehouse, where we have previously extracted hundreds of features from our users as of January 2015. In this research, we perform two-group chi-squared tests and ANOVA of the analysis framework to find features that have statistically significant differences between the positive samples (heavy users) and negative samples (non-heavy users). Specifically, we refer to features that have very high η^2 values in chi-squared tests and very high F-statistic values in ANOVA.

The connection records collected in Chunghwa Telecom can be categorized in two major sources: mobile and (northern Taiwan) broadband. In the following sections, we define heavy users of top video apps in Section V.A, and show the profiling results for mobile and broadband users respectively in Section V.B and Section V.C:

A. *Definitions of Heavy Users*

In this research, we use usage days to define heavy users of the top video apps. Although video access time would be a much more accurate measurement, the connection records collected from our network devices are merely sampled data. Thus, we believe usage days are a more credible measurement.

Within the 8-week period of our observations, the distribution of the number of days for which a user have used *any* of the top video apps is shown in Fig. 2. For easier presentations, we exclude the users who use 0 day of the apps.

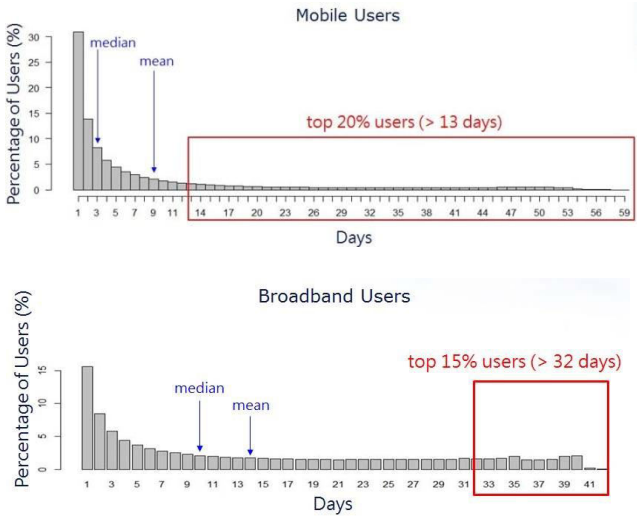

Figure 2. Distribution of usage days of the top video apps.

For both mobile and broadband users, we can observe a roughly power-law distribution, i.e., the percentage of users decay sharply in the first few number of days, and has a long tail beyond the mean. The mean and median for mobile users are 9.0 days and 3 days, and the mean and median for broadband users are 14.4 days and 10 days.

To meet the needs of the marketing division, in this research, we define "heavy users" as those mobile users whose usage of the top video apps is in the top 20% (13 days and above), and those broadband users whose usage is in the top 15% (32 days and above). We choose these percentages to control the number of users for our marketing plans.

B. Profiling Results of Mobile Heavy Users

For mobile users, the distributions of gender and age are shown in Fig. 3 and Fig. 4. In terms of gender, the ratio of females to males is 1.24 among heavy users, whereas among non-heavy users, this ratio is only 0.87. In terms of age, the average of heavy users (37.9 years old) is younger than non-heavy users (45.1 years old), and there is a much higher percentage in ages 18 through 24 among heavy users than among non-heavy users. This matches our intuition that females and the youth age group would account for higher proportions among video app users.

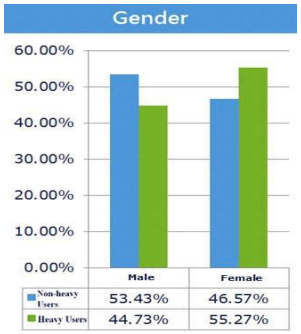

Figure 3. Gender distribution among heavy and non-heavy users.

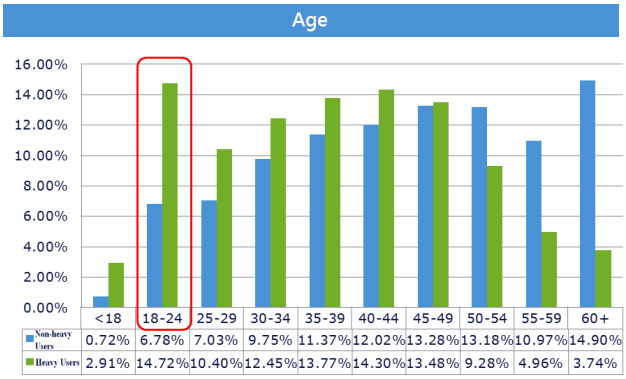

Figure 4. Age distribution among heavy and non-heavy users.

In terms of moving distances and phone call patterns, Table I. shows that heavy users are more inclined to travel around and have more contacts to call. For example, on weekdays, heavy users travel 151 kilometers more than non-heavy users; on weekends, they travel 55 kilometers more. It is possible that heavy users are often engaged in lengthy commuting, and they often watch videos on the train, bus, or subway. In addition, heavy users have more cities of callers, more base stations of callers, longer total length of call durations, more contacts called to, and higher proportion of off-net subjects (whose carrier is not Chunghwa Telecom) called to.

TABLE I. MOVING DISTANCES AND PHONE CALL PATTERNS.

Feature name	Avg. of non-heavy users	Avg. of heavy users
Travel distance on weekdays (km)	364.7	516.3
Travel distance on weekends (km)	128.4	183.5
# of cities of callers	3.8	4.5
# of base stations of callers	15.8	19.9
Total length of call durations (minutes)	90.4	123.6
# of contacts called to	15.7	18.3
Proportion of off-net subjects called to	25.3	28.8

The significant mobile features of heavy users are also observed when purchasing a new cell phone. In this research, heavy users spend 49% more money when buying the last cell phone from our company and their purchase frequency is 23% higher.

Finally, in terms of mobile plans, users subscribing to flat rates have a 13.8% probability of being heavy users, much higher than a 5.7% probability when subscribing to non-flat rates. Also, heavy users pay 70% more mobile fees than non-heavy users, measured in the last 3 months' averages.

C. *Profiling Results of Broadband Heavy Users*

Heavy users might have the same interests on the internet activities. In order to distinguish which users have the same interest, we classify and label website by web crawling. Then we calculate the vertical interest scores and horizontal interest scores for each user by formulas.

The vertical interest score is to calculate i customer interest score based on interest j and compare with other users. The horizontal interest score is to calculate user j on all interests and compare with other users. By the following formulas, we calculate each customer's vertical interest score and horizontal interest score.

$Int\_V_{i,j}$ means customers i with interest j in vertical interest scores, $Int\_H_{i,j}$ means customers i with interest j in horizontal interest scores.

$$Int\_V_{i,j} = \sqrt{\left(\frac{C_{i,j}}{Max_j(C)}\right)^2 + \left(\frac{D_{i,j}}{Max_j(D)}\right)^2} \quad (1)$$

$$Int\_H_{i,j} = \sqrt{\left(\frac{C_{i,j}}{Max_i(C)}\right)^2 + \left(\frac{D_{i,j}}{Max_i(D)}\right)^2} \quad (2)$$

C: Frequency of use per month

$C_{i,j}$: Frequency of use per month in customers i and interest j

D: Number of use days per month

$D_{i,j}$: Number of use days per month in customers i and interest j

For example, suppose the records of browsing all categories for both customer A and B are 30 days. We calculate social category interest scores as shown in Table II.

TABLE II. INTEREST SCORE

HN	social category	video category	horizontal times (Max)
customer A	10	20	20
customer B	5	0	5
vertical times (Max)	10	20	

$Int\_V_{i,j}$, vertical interest scores, is calculated by formula (1). $Int\_H_{i,j}$, horizontal interest scores, is calculated by formula (2).

$$Int\_V_{i,j} = \sqrt{\left(\frac{10}{10}\right)^2 + \left(\frac{30}{30}\right)^2} \approx 1.414$$

$$Int\_H_{i,j} = \sqrt{\left(\frac{10}{20}\right)^2 + \left(\frac{30}{30}\right)^2} \approx 1.118$$

The interest attributes analysis performed by the above formula show that heavy users have a variety of interest, especially about video, e-commerce, gaming and community activities. Their life highly relies on the network to deal with almost everything, as shown in Table III.

TABLE III. INTEREST IN FEATURE LIST

Feature name	Avg. of non-heavy users	Avg. of heavy users
[DNS Interest] video vertical of interest percentile	46.643	68.367
[DNS Interest] e-business vertical of interest percentile	44.665	65.76
[DNS Interest] game vertical of interest percentile	47.549	65.533
[DNS Interest] social vertical of interest percentile	48.251	64.494
[DNS Interest] news vertical of interest percentile	46.005	60.863

In terms of business combinations of customer used, heavy users prefer popular video service. For Chunghwa Telecom current mobile users, the mobile phone marketing programs and the contracts with video service added will be suitable for them. For non-Chunghwa Telecom current mobile users, the mobile number portability program with video services could be considered. By providing more suitable business combinations service, users may have higher will to use their video services.

In terms of Internet bandwidth, users subscribing to high speed Internet have a higher probability of being heavy users. Also, heavy users download more data, about 1.5 times that of non-heavy users.

VI. EVALUATION

In order to evaluate the heavy user's profiling accuracy, this research provide the analysis results to the business owner marketing unit, to help them understand the heavy user's profiles and to design suitable activities and useful promotion plans. Our company hopes these new methods will make a good impression on customers and attract them to adopt their services.

Our company can send eDMs, using their CRM (customer relationship management) promotion platform, to these target customersŁemail boxes which they often used. The eDMs can provide customers about top HOT films information and the best charge scheme. On the other hand, the company can post these great information on video-related forums and fans group to promote their service.

The research provides a new way to use the Chunghwa Telecom big data analysis platform for big data analysis, management and calculation. Applying this method for promotion activities is still in progress. Currently, the high-usage users profiling analysis on mobile and broadband had been completed. The analyzed results had been provided to the unit of business owner to make suitable activities and marketing plans. We expect the number of paid customers could be increased.

The other successful using case of the platform is the one completed in 2014 called "Chunghwa Telecom Public Wi-Fi site selection analysis". The case needed to decide the Wi-Fi site locations via mobile voice and data access users

locations, GPS positioning and Wi-Fi hotspot. Finally, the potential Wi-Fi hotspot locations were provided by the platform to unit of business owner for further development. Analysis results show that the average throughput of those 50 suggested hotspots is higher than others by about 15%.

The "Broadband loss warning analysis", completed in 2013, is another successful case. The case analyzed broadband DNS browsing log information to find the customers who may cancel services. The company adopted suitable customer care program to increase their loyalty. Analysis results shown that customer loss rate is lower than the average value by 12%.

VII. CONCLUSIONS

In this research, we have used big data computations in the network connection records in Chunghwa Telecom to define “heavy users” of 12 top video apps. In addition, we have exploited hundreds of user features in our data warehouse to run chi-squared and ANOVA tests to find features that have statistically significant differences between heavy and non-heavy users. Such profiling results can be used to design corresponding marketing plans.

Discovered in our profiling results, heavy users on average have a higher proportion of females, younger ages, longer travel distances, more extensive phone call patterns, higher expenses and higher frequencies of cell phones purchases, and higher mobile fees, higher broadband speeds, and higher download volumes. Also, they have higher interest scores in videos, e-commerce, games, and social media.

Currently, our research only performs user profiling of top video apps. In the future work, we will use the analytic techniques introduced in the analysis framework in Section IV to perform accurate predictions. For example, we can use classification algorithms to predict potential users of our own video services, toward the goal of target marketing.

REFERENCES

[1] W.T. JHENG, "Cross-business product sales method." TW 201120778, JUN. 16, 2011.
[2] N.Y. JAN, "Data mining model system and method for automated maintenance and operation" TW 201426352, JUL. 1, 2014.
[3] R. Bambini, P. Cremonesi, and R. Turrin, A Recommender System for an IPTV Service Provider: a Real Large-Scale Production Environment recommender Systems Handbook, Springer, 2010, ch.9.
[4] H.C. Yeh and Z.F. Wu, “A Study on the Application of Data Mining on Mobile Value-Added Marketing - Based on the Telecommunication Kaohsiung Area Users,” 2013, pp.14–21.
[5] Y.X. Li, “Cross-selling of data mining applications in the telecommunications company,” 2014, pp.24–38.
[6] P. H. Swain and H. Hauska, “The decision tree classifier: Design and potential,” IEEE Trans. Geosc. Electron, vol. 15, no. 3, July 1977, pp. 142–147.

About Authors

[1] C Sarishma,
PG Scholar, Dept of MCA,
SIETK, PUTTUR.

[2] P Balaji,
Asst. Professor, Dept of MCA,
SIETK, PUTTUR.