Insights from Analysis of Video Streaming Data to Improve Resource Management

Sabidur Rahman*, Hyunsu Mun†, Hyongjin Lee‡, Youngseok Lee†, Massimo Tornatore*, and Biswanath Mukherjee*
*University of California, Davis, USA †Chungnam National University, Korea ‡Politecnico di Milano, Italy
Email: {krahman, mtornatore, bmukherjee}@ucdavis.edu, {munhyunsu, hjlee201203399, lee}@cnu.ac.kr

Abstract—Today a large portion of Internet traffic is video. Over The Top (OTT) service providers offer video streaming services by creating a large distributed cloud network on top of a physical infrastructure owned by multiple entities. Our study explores insights from video streaming activity by analyzing data collected from Korea’s largest OTT service provider. Our analysis of nationwide data shows interesting characteristics of video streaming such as correlation between user profile information (e.g., age, sex) and viewing habits, viewing habits of users (when do the users watch? using which devices?), viewing patterns (early leaving viewer vs. steady viewer), etc. Video on Demand (VoD) streaming involves costly (and often limited) compute, storage, and network resources. Findings from our study will be beneficial for OTTs, Content Delivery Networks (CDNs), Internet Service Providers (ISPs), and Carrier Network Operators, to improve their resource allocation and management techniques.

Index Terms—User behavior; video streaming; cloud network management; data analysis; Apache Spark.

I. INTRODUCTION

Internet users are using more and more Video on Demand (VoD) applications such as Netflix, Amazon, YouTube, etc. Even traditional live television channels are now being streamed by service providers such as Sling TV, DirectTV Now, etc. According to Cisco Visual Networking Index (VNI) [1] 80-90% of the traffic in 2021 will be video.

VoD services require compute, storage, and network resources which are costly and often limited. Hence, new effective resource management methods are required to serve ever increasing video traffic. These methods should be scalable, adaptive, and aware of patterns related to user activity, spatio-temporal variation of load, user profiles, devices used, etc. Hence, understanding these patterns is very important for resource management of carrier networks, DC networks, Content Delivery Networks (CDNs), Internet Service Providers (ISPs), etc.

There have been a few prior works studying partial analysis of VoD services. For example, Ref. [2] analyzes user behavior in VoD traffic, specifically video watching sessions. The data analyzed in the study was collected before 2014 and uses Hadoop system to process the data. Ref. [3] investigates characteristics of user behavior in mobile live streaming systems. But the data only accounts for live contents and mobile devices, hence it misses out on large portion of contents and users. Due to privacy concerns, limitation of publicly available data from large video streaming services, this area of study lags behind compared to areas such as analysis of mobile traffic patterns [4] and mobile application usage [5]. To fill this void, our study reports insights from a comprehensive nation-wide dataset including live channels and video contents, traces collected over both wireless and wired network, etc. Another unique feature of the data is that it not only contains user’s viewing traces, content types and identification in details, but also contains anonymized information such as user’s age group, device IP address, device type, etc.

To the best of our knowledge, our study is the first to use big data analytics tools Apache Spark [6] and Zeppelin [7] to analyze video streaming activity. The size and scale of data used in our study required more than traditional analytics tools. To keep the analysis scalable for future studies and deployable in practical run-time scenarios, we have used big data analytics tools. The results from our analysis enable us to explore answers for the following questions:

- Analysis of user profiles: Who is watching the videos?
- Temporal patterns in viewing activity: When do users watch video?
- Content-centric viewing patterns: Which contents are popular? When are popular contents viewed?
- Video browsing behavior: How can we correctly classify video browsing behaviors (early leaving, steady viewer, etc.)?
- User device: Which device(s) do users use to watch certain videos?

The rest of this study is organized as follows. Section II provides a description of the dataset and methodology. Section III describes how we use big-data analytics to answer questions that are important for infrastructure owners and service providers. Section IV concludes the study and indicates directions for future works.

II. DATASET AND METHODOLOGY

A. Description of Dataset

Our study uses dataset collected from the largest OTT in Korea. This nationwide data was collected in 2017 and the results in our study are derived from 24-hour data collected from 3M subscribers on a weekday.

The data considers 70 live channels, 7000 movies, and 280,000 other VoD contents. The data trace was collected every 10 seconds. Tables I and II show sample data with explanation followed.
TABLE I: Sample data: viewing information.

| D | H | M | S | U. ID | T. ID | P. ID | M. T. | Dev. | BR | IP |
|---|---|---|---|------|------|------|------|------|----|----|
| d10 | 01 | 10 | a3 | L | A01 | 01:20 | Andr. Ph. | 2 | p | p |
| d10 | 01 | 20 | a3 | L | A01 | 01:30 | Andr. Ph. | 2 | p | p |
| d10 | 01 | 20 | a5 | V | A01 | 00:05 | Smrt. TV | 1 | q | q |

In Table I, the first four columns (D, H, M, and S) display date, hour, minute, seconds of the trace. The fifth column contains user identification (e.g., a3) which helps us to map the viewing info with user profile from Table II. Sixth column (T) reports ‘the type of video content’: ‘L’ for live contents, ‘V’ for non-live video contents. Seventh column (P.ID) contains the content identification. Eighth column (M.T.) contains the media time inside the content where the viewer is watching now. Ninth column (Dev.) contains the device type the viewer is using. We have seven major types of devices in our study data: Android Phone, iOS Phone, Android Tablet, iOS Tablet, PC, Smart TV, Chromecast. Tenth column (BR) contains the ‘bitrate’ requirement for the connection (‘2’ stands for 2 Kbps and so on). Eleventh column (IP), contains the IP address (e.g., ‘p’) of the customer device with lower 16 bits hidden (for example, ‘p’ = 192.168.*.*).

Table II contains user profile information such as gender and age group of the user. This helps us to explore the impact of such information in viewing behavior and more.

TABLE II: Sample data: user profile information.

| User ID | Gender | Age Group |
|---------|--------|-----------|
| a3      | M      | 30        |
| a5      | F      | 20        |

B. Methodology and Analysis Environment

For big data analysis environment, we have used Apache Spark 2.2.0 and Apache Zeppelin 0.7.3. Apache Spark is an advanced analytics engine for large-scale data processing. Apache Spark uses both batch and streaming data to gain high performance. Ref. [6] reports hundred times faster performance compared to Hadoop.

Traditional methods such as standard database queries and Hadoop distributed processing can generate these results as well, when there is no computation time constraints. But, when we start applying this results in real-time, tools like Apache Spark are a better option in terms of computation time and large-scale data handling.

III. FINDINGS AND RESULTS: WHAT QUESTIONS CAN DATA ANSWER?

We study and analyze the data to understand the relationship between user activity and features such as time of the day, age group, gender, device type, content type, etc. From our findings through data analysis, we report the following results.

A. Analysis of user profiles: Who is watching the videos?

Table III shows the breakdown of number of viewers (%) in different age groups. According to Table III, the maximum number of users come from age group 30 (39.54%), followed by age group 20 (31.09%). Such findings can help to understand the distribution of users’ age groups.

TABLE III: Age group analysis for viewers.

| Age Group | Percentage |
|-----------|------------|
| Less than 10 | 0.001 |
| 10-19 | 2.48 |
| 20-29 | 31.09 |
| 30-39 | 39.54 |
| 40-49 | 17.16 |
| 50-59 | 7.09 |
| 60-69 | 1.78 |
| Greater than 70 | 0.798 |

The gender breakdown of viewers is 53.54% female, 42.68% male and 3.78% others.

B. Temporal patterns in viewing: When do users watch video?

1) Impact of age: Fig. 1 shows the actual number of users over 24 hours, distributed among different age groups. We observe that viewing activity increases as the day progresses, for all age groups. Also, viewing activity increases rapidly after 1900, one explanation of this can be: as more users come back home, number of active viewers grows. In addition, users from age groups 20 and 30 dominate throughout the day. This phenomenon can be explained by Table III, as we have observed the largest number of viewers come from age group 20 and 30.

Fig. 1: Actual number of active viewers over 24 hours, from different age groups.

2) Impact of gender: Fig. 2 shows the hourly user count over 24 hours, distributed among different genders. We observe similar patterns as Fig. 1. Viewing activity increases as the day progresses for both male and females. Contrary to popular beliefs, we observe more female viewers throughout the day. This phenomenon, again, can be explained by the fact that larger portion of the users are from female gender.

C. Content-centric viewing patterns: Which contents are popular? When are popular contents viewed?

Fig. 3 shows very interesting pattern in three different types of contents. Content relevant for “Kids” is popular most of the active day-time and late at night. On the other
hand, news related content is popular only at night (1900-2300) as the adult age group are actively viewing during that period. Similar pattern is also followed by “Sports”. Content placement and network management methods unaware of such trends might make inaccurate decisions, leading to QoS violations and additional operational cost.

D. Video browsing behavior: How can we correctly classify video browsing behaviors?

One of the most interesting insights of our study is the viewer’s behavior during the viewing. After studying the data, we have classified the viewer’s viewing pattern in the following four categories:

- Early leaving: stops watching content in 5 minutes. 
- Steady viewer: watches a single video for more than 5 minutes.
- Highlighter: browses only the interesting parts of a video.
- Surfing watcher: watches several videos within an hour.

Our data indicates that the viewers do not usually watch the whole content. A large portion of them watches only the interesting parts (highlighters are 56.44%), followed by the steady viewers (25.74%). In addition, there is a significant number of early leaving viewers (10.89%) too, who leaves the system in less than 5 minutes. Surfing viewers, who watch multiple contents within an hour consists of 7.0%. Hence, resource management methods should prepare the system to serve highlighters and surfing watchers, in addition to steady and early leaving viewers.

E. User device: Which device(s) do users use to watch videos?

From our analysis, the breakdown of the devices used for video streaming is as follows: ‘Android Phone’ leads with 33.5%, followed by ‘PC’ 29.5% and ‘iOS Phone’ 16.6%. Rest of the devices are ‘iOS Tablet’ (9.6%), ‘Android Tablet’ (5.1%), ‘Smart TV’ (5%), and others (0.7%). Content quality and content size (4K HD vs. 1080P) will change with the device and screen size making the device type is an important parameter in content and network management problems.

IV. Conclusion

Our study presents an analysis of nationwide data collected from Korea’s biggest OTT service provider. We use big data analytics tools (Apache Spark and Zeppelin) to analyze the data. Our analysis shows interesting insights into user behavior and helps us to look for answers of important questions related to video streaming. Future studies can use such understandings to develop network and content management methods which are more data-driven. Predicting future user behavior (steady viewer vs. early leaving) by learning from previous patterns is one interesting direction to explore.

Acknowledgement

Youngseok Lee was supported by Basic Science Research Program through the NRF of Korea funded by the Ministry of Education (NRF-2016R1D1A1A09916326) and by the MSIT of Korea under the ITRC support program (IITP-2018-2016-0-00304).

REFERENCES

[1] “The Zettabyte Era: Trends and Analysis”, Cisco, [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html. [Accessed: April 06, 2018]
[2] L. Chen, Y. Zhou, and D. M. Chiu. “A study of user behavior in online VoD services,” Computer Communications, vol. 46, pp. 66-75, 2014.
[3] Z. Li, G. Xie, M. A. Kaafar, and K. Salamatian, “Characterizing and Modeling User Behavior in a Large-scale Mobile Live Streaming System,” IEEE Trans. on Circuits and Sys. for Video Tech., 2016
[4] F. Xu, Y. Li, H. Wang, P. Zhang, and D. Jin. “Understanding mobile traffic patterns of large scale cellular towers in urban environment,” IEEE/ACM Trans. on Networking, vol. 25, no. 2, pp. 1147-61, 2017.
[5] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, C. Ziemlicki, Z. Smoreda. “Not All Apps Are Created Equal: Analysis of Spatiotemporal Heterogeneity in Nationwide Mobile Service Usage,” Proc. of 13th International ACM Conference on Emerging Networking Experiments and Technologies, 2017.
[6] “Apache Spark”, Apache, [Online]. Available: https://spark.apache.org/ [Accessed: March 12, 2018]
[7] “Apache Zeppelin”, Apache, [Online]. Available: https://zeppelin.apache.org/ [Accessed: March 12, 2018]

Prior study \[2\] examines the early-leaving viewer behavior and shows why ‘5 minutes’ is a practical threshold.