Cellular Smartphone Traffic and User Behavior Analysis

Yinzhou Li, Jie Yang
School of Information and Communication Engineering
Beijing University of Posts and Telecommunications
Beijing, China, 100876.
Email: arenalyz@gmail.com

Nirwan Ansari
Electrical and Computer Engineering Department
New Jersey Institute of Technology
New Jersey, US, 07102.
Email: nirwan.ansari@njit.edu

***Abstract*—Recent emergence of smartphone applications have led to explosive traffic growth in cellular networks. Understanding the traffic characteristics and user behaviors in cellular data networks becomes critical in the rapidly evolving market. Statistics show that Android and iOS are two leading smartphone operating systems and Windows Phone operating system is catching up fast in the global smartphone market share. This paper characterizes mobile Internet traffic generated by Android, iOS, and Windows Phone platforms devices. We also explore and compare user behaviors of these three platforms from two aspects: traffic dynamics and user applications. This study was conducted based on the traffic data collected from a major cellular operator's network covering more than two million users. The platform of a mobile device is identified based on HTTP signatures. Our analysis of this big data set is crucial for improving user experience and future smartphone application design.**

I. Introduction

Cellular data traffic has been growing dramatically. According to a recent study, global mobile data traffic will continue to grow. Furthermore, smartphone traffic contributes 92% of the total global handset traffics in 2012 [1]. The 'open' environment of smartphone platforms has spurred a great diversity of applications (*apps*). The iOS AppStore, Android Google Play, and Windows Mobile MarketPlace enable users to acquire their apps easily. Apple officially announced that customers had downloaded over 50 billion apps from AppStore by May 16, 2013 [2]. To help cellular network operators design more appropriate networks and application developers create applications which could better serve user needs, the patterns of Internet traffic carried by cellular networks should be understood and the user behavior of major smartphone platforms should be characterized.

This paper studies the traffic pattern and user behavior on three major smartphone operating systems (*OS*): Android, iOS and Windows Phone. We measure massive flow traces collected from the network of a major service provider in China. The big data set we used is in the magnitude of Terabytes, which are saved in Hadoop Distributed File System (*HDFS*). HDFS is designed to store super large files across machines in a distributive manner and measure the data via the MapReduce (*MR*) framework [3]. The scale of the dataset in our study is one of our key contributions: it is at least an order of three, in terms of user sets, larger than a prior work [4], which investigated the diversity between Android and Windows Phone based on 255 user sets. Moreover, our study includes an additional platform, i.e., Apple iOS.

User behavior patterns are highly correlated with device features and capabilities, which are in turn limited by the device hardware and platform. Therefore, understanding different platform user behaviors can help OS vendors design more efficient platforms and also help network operators achieve better user experience.

The rest of the paper is organized as follows. Section II summarizes related works on cellular networks traffic analysis and mobile device recognition. Section III describes our traffic traces and our method in identifying device platforms. Section IV presents overall temporal traffic patterns of three smartphone platforms. Section V illustrates detailed user behavior analysis of the three platforms in terms of traffic dynamics and user applications. Conclusion is drawn in Section VI.

II. Related Works

Recent studies have shed valuable insights on some aspects of smartphone traffic analysis. Maier *et al.* [5] gave an insight into the smartphone traffic from transmission perspective and found out that browsing contributes over half of the traffic. Shafiq *et al.* [6] studied the cellular traffic dynamics from the perspectives of the device type behavior and the application behavior. They analyzed the dataset of two smartphone device types and a family of cellular broadband modems. They also provided a model to predict future traffic patterns. Alessio *et al.* proposed a service condition concept to improve the accuracy of measuring and analyzing IP networks [7]. With this concept, they evaluated the network performance by taking account of devices and operating systems as well as other aspects. In [8], Ram *et al.* investigated whether there exist distinct behavior patterns among 3G mobile network users using clustering method. Results showed that the browsing behavior patterns of mobile users can be classified by a small number of co-clusters, and exhibit stability at short and long time scales. Similar to our study, Qiang *et al.* [9] presented results of the diverse usage behaviors of smartphone users from mobile application perspective. In contrast to previous work, we focus on characterizing user behaviors of three

978-1-4799-2003-7/14/$31.00 ©2014 IEEE

leading OS platforms of smartphones by analyzing massive traffic records in a large tier-1 cellular network in China.

There are standard ways [10][11] to retrieve device capabilities from User-Agent (*UA*) Request Headers [12] and Accepted Request Headers encapsulated in HTTP requests. However, many mobile devices do not conform to the specifications of standard ways. Wireless Universal Resource File (*WURFL*) [13] addresses this issue by adopting a UA-based device recognition method, which enables web servers to recognize information of an accessing device by matching the UA content with a predefined configuration file. The configuration file contains information such as model, brand and platform of more than eighteen thousands of device models. For the same purpose of recognition device model as WURFL, a method based on Jaccard measurement is proposed in [14], which can identify a proper keyword of mobile device model. In this paper, we identify smartphone platforms by matching platform keywords in the WURFL configuration file, and use the 'same model same operating system' criterion to enhance the platform identification rate.

III. Data Set

A. Data Collection

The big data set used in this paper was captured from a Universal Mobile Telecommunication System (*UMTS*) network [15] of a leading cellular network operator in China. We collected the data from all links between Serving PRS Support Nodes (*SGSNs*) and Gateway GPRS Support Nodes (*GGSNs*) in a southern city of China. In the UMTS network, a mobile device forwards its voice or data traffic to the Radio Network Controller (*RNC*) through a cell tower (node-B) and the RNC connects to the core network via SGSN and GGSN nodes. GGSNs are responsible for providing connectivity to external networks. Hence, the data set, collected by our Traffic Monitoring System (*TMS*), contains all information carried in the Packet Data Protocol (*PDP*), including the app identifier, viz. the TMS app identifier, which classifies the traffic into different categories and applications by applying DPI (Deep Packet Inspection) and DFI (Deep Flow Inspection) [16] methods. A category refers to a general class grouped by its usage. For example, P2P download refers to the applications which offer user downloading resources from network via peer-to-peer communications. Fig. 1 illustrates the architecture of the cellular network used for this study and the deployed TMS.

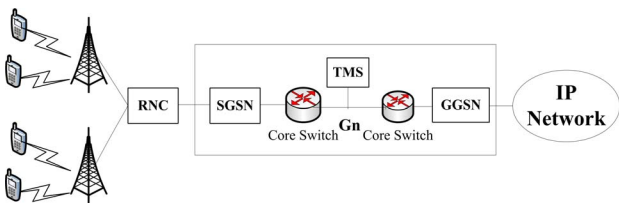


Fig. 1. UMTS network architecture.

The data set generated by TMS is stored in the format of flow logs. All information in flow logs is based on the flow level. Each record in flow logs contains flow connection time, flow upload and download bytes, International Mobile Equipment Identifier (*IMEI*) number [17], categories and applications, user-agent field (HTTP requests header) if the flow is carried by the http protocol, etc. Each IMEI represents an individual mobile device. We anonymize device IMEI numbers to protect the privacy of subscribers.

B. Data Measurement

In this paper, we captured two whole days of UMTS network activities on May 5th 2013, Sunday and Feb. 20th 2013, Wednesday, respectively. The flow logs from TMS contain a massive number of IP accessing records including HTTP flow records. To analyze the traffic pattern and user behavior of different operating systems, users of each OS should be recognized beforehand. We first abstract OS keywords and browser keywords from WURFL. Here, keywords are inscribed in the OS field of UA. Including the operating systems and browsers commonly used in China uncovered by WURFL, a total number of 29 kinds of browsers are identifiable. Based on these keywords in UA, the OS can be recognized according to the following procedure:

1) (step 1) OS Recognition: UA carries rich information about mobile device characteristics, but many UA headers generated by applications are not compatible with those originated by browsers. Thus, we only employ the http flow records originated by browsers to recognize the device OS. If the UA contains browser keywords, it is considered as originated by a browser upon which further match with OS keywords will be conducted. We list the IMEI number and UA in each http flow record originated by browsers in a table. In this table, one IMEI number can correspond to a number of UAs because one handheld device usually generates more than one flow in online activities. We match each UA with OS keywords and calculate the number of UAs grouped by the OS keywords. For one IMEI number, the OS is identified as that with the majority of the corresponding OS keyword appeared in UAs. For example in Fig. 2, IMEI1 has three UAs. Two of them contain OS1 keyword, and one contains OS2 keyword, and so the OS of IMEI1 is OS1.

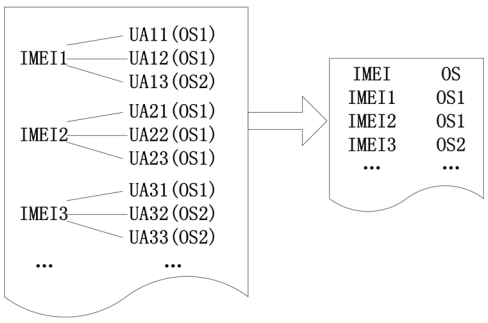


Fig. 2. OS recognition by keyword matching.

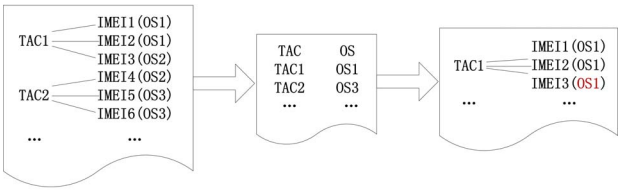


Fig. 3. Enhanced OS recognition.

2) (step 2) Enhanced OS recognition by device model: The Type Allocation Code (*TAC*) number, which can be retrieved from the first eight digits of IMEI, identifies the manufacturer and model of the device. Therefore, there are several IMEI numbers under each TAC lot. Normally, one device model is, by default, to run one specific OS when produced by the manufacturer, and so all the IMEI numbers belonging to one TAC correspond to the same OS. In the last step, the OS of each IMEI number is identified by keyword matching. To accurately recognize the OS of each IMEI, we use the TAC number to revise device platform information. Each TAC lot contains many IMEI numbers. Each IMEI number is associated with an OS based on the result of step 1. The OS associated with the most number of IMEIs is considered as the OS of this TAC. Then, the OS of all the IMEIs in the TAC lot would be updated with the same TAC OS. Thus, the recognition of IMEI OS is enhanced. For the example shown in Fig. 3, TAC1 has three IMEI numbers; after step 1, two of them are tagged with OS1, one is tagged with OS2, and so the OS of TAC1 is OS1. Then, the OS of IMEI3 is revised to OS1.

IV. Traffic Volume Analysis

Table I summarizes our collected big data set in terms of the numbers of recognized users and traffic volumes that employ different operating systems. Data set 1 represents the flow information of May 5th 2013 and data set 2 represents that of Feb 20th 2013. Here, we focus on characterizing the traffic volume and user behavior based on the data collected on May 5th 2013. The data set on Feb 20th 2013 will be used in the next section to verify our findings.

Table II shows the variance of the normalized traffic volume of the three OS traffics at per hour granularity in one day. Note that we normalize the traffic volume of each hour by the total traffic volume of its device OS family. The results show that traffic of Windows Phone OS exhibits relatively high variance, implying that Windows Phone users tend to use their cellular devices more intermittently.

In order to mine the detailed traffic characteristics among these three operating systems, we divide 24 hours into 4 periods of time. Each period covers 6 hours. Table III shows the traffic volume percentage of each individual period with respect to the overall traffic for each OS. The results show that a quarter of a day at night contributes one third of the total traffic in a day.

V. Measuring User Behavior Dynamics

In this section, we investigate how each OS device user uses the cellular network from traffic and applications perspectives. Here, we first analyze the traffic diversity of each individual of the three OSs. Then, we compare user behavior over app in two perspectives among these three OSs. One is app category usage and the other is app diversity of each user used in one day.

A. User traffic dynamics

In the previous section, we depict the characteristics of overall traffic of each OS. The result shows that Windows Phone OS traffic fluctuates the most. In this section, we analyze traffic dynamics from the user perspective. Table IV shows the mean value of user traffic throughout the day. It is obvious that iOS platform users generate much more traffic than the other two OS users. To learn more about user traffic dynamics of each platform, we group users by traffic volume characteristic as high/medium/low traffic (*HT/MT/LT*). The traffic volume is normalized by the maximum observed value for each user. If the average normalized traffic for each 5-minute is more than 0.5, this user traffic is considered as high traffic. If the average normalized traffic is no more than 0.5 and no less than 0.1,this user traffic is considered as medium traffic. Finally, if the average normalized volume is less than 0.1, it is categorized as low traffic. Then, we calculate the user percentage of each group in each platform. The results shown in Table IV indicate that traffic generated by most of users in each OS family belongs to the LT group, implying that most users have only a few obvious traffic peaks and considerably low traffic volume in most of the time.

TABLE I
DATASET SUMMARY

OS	Number of Users		Traffic Volume (GB)	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2
Android	1,725,092	1,721,411	7543.76	7396.01
iOS	11,999	9,289	164.40	214.77
Windows Phone	17,180	11,815	78.25	56.79

TABLE II
VARIANCE OF TRAFFIC

OS	Android	iOS	Windows Phone
Variance	1.198	1.606	1.954

TABLE III
TRAFFIC PERCENTAGE OF EACH TIME PERIOD

OS	1:00-6:00	7:00-12:00	13:00-18:00	19:00-24:00
Android	19.16%	26.17%	23.63%	31.04%
iOS	15.93%	25.17%	27.92%	30.98%
Windows Phone	15.75%	26.84%	26.50%	30.91%

TABLE IV
TRAFFIC AND USER RATIO OF TRAFFIC GROUPS

OS	Mean Traffic	HT	MT	LT
Android	4.37 MB	11.25%	15.63%	73.11%
iOS	13.70 MB	5.52%	9.40%	85.08%
Windows Phone	4.55 MB	8.36%	10.63%	81.01%

B. User application dynamics

According to the above traffic volume analysis, we are investigating traffic dynamics from the application perspective. Using application categories and names labeled by TMS, we conduct our study from both categories and applications aspects.

1) User Percentage of Application Categories: Here, we define three OSs as $O = \{o_1, o_2, o_3\}$, which denotes Android, iOS and Windows Phone, respectively. All application categories are represented as $C = \{c_1, c_2, ..., c_N\}$. We have aggregated accessing records r for each user. Each record $r = \langle u(r_i), o(r_i), c(r_i)\rangle$ represents that a user used a device with identification $u(r_i)$ employing OS $o(r_i)$ to access the application $c(r_i)$ at least once, where $o(r_i) \in O$ and $c(r_i) \in C$. Then, the user ratio of app category c_j in o_i, defined as $P(o_i, c_j)$, can be computed according to Equation 1.

$$P(o_i, c_j) = \frac{H_{o_i c_j}}{H_{o_i}} \tag{1}$$

where $H_{o_i c_j}$ represents the number of users whose device OS is o_i and who have used application category c_j, and H_{o_i} is the number of users whose device OS is o_i.

Figure 4 shows the ratio of users of 7 major application categories including browser, IM, game, weibo, download, SNS and appMarket, in each OS on May 5th 2013. More in detail, browser refers to the web browser applications like Safari and Chrome. IM refers to the applications which offer instant message service, such as Microsoft messenger. Game refers to the game applications running on mobile devices. Weibo is a particular kind of applications in China which are similar to Twitter. SNS is short for Social Network Site, including a number of websites such as LinkedIn and Facebook; AppMarket refers to the applications which offer app downloading for users like Apple AppStore. in each OS on May 5th. As compared to other application categories, browsing is the most active application with the largest number of users. It is reasonable as browsing is the most basic online behavior. iOS users exhibit the biggest user ratio towards appMarket category, implying that iOS users are fond of downloading apps from apple Market. For the download category, Windows Phone yields the biggest user ratio. To verify our findings in the usage of application categories, we conduct the same analysis on data set 2, which is a weekday. The result is shown in Fig. 5. Again, browser is used by most users in all three platforms, and appMarket continues to be the most popular application in iOS as compared to the other two platforms. However, Windows Phone does not yield the biggest user ratio in the download category, but still maintains a considerably high value.

Figures 6 and 7 present the traffic volume generated by each user in individual application categories. Fig. 6 shows the result of Sunday, and Fig. 7 represents the result of Wednesday. They show that browser and download incur the

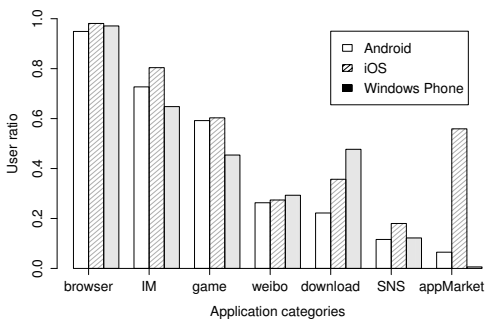


Fig. 4. User ratio of each application category on May 5th 2013.

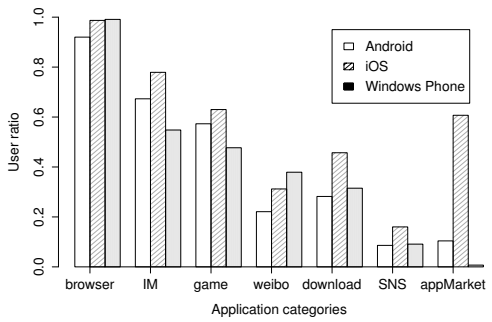


Fig. 5. User ratio of each application category on Feb 20th 2013.

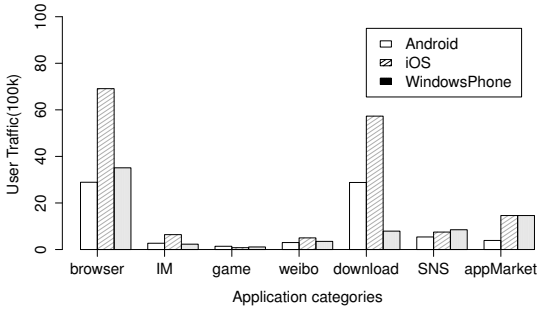


Fig. 6. The traffic volume of each category per user on May 5th 2013.

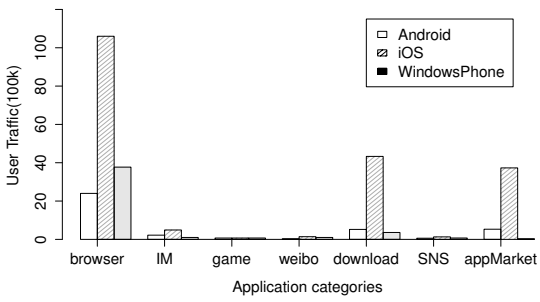
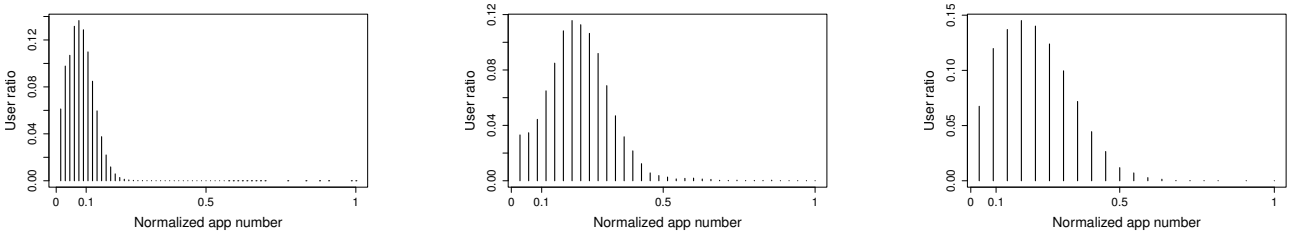


Fig. 7. The traffic volume of each category per user on Feb 20th 2013.



(a) (b) (c)

Fig. 8. User distribution of app diversity: (a) Android, (b) iOS, and (c) Windows Phone.

major traffic volume of all three OS users both on weekday and weekend. iOS users generate a larger appMarket traffic volume on weekday than on weekend. Note that only a very small group of Windows Phone users have used appMarket and also generated mere traffic volume per user. This is attributed to the fact that Windows Phone app market offers less app as compared to AppleMarket and the open Android market. It is an interesting phenomenon that the user ratio of micro blog (an online social networking service like twitter in China) and SNS do not change much between weekday and weekend, but the traffic of these two categories on weekday are much less than on weekend, implying that people always keep their SNS pages online, but seldom update them on weekdays.

2) The Number of Apps per User: Normally, a user would not use only one service with his/her mobile device in a day. The number of apps a user used in one day depends on several factors, such as his/her job, personality and also the device model. Each OS is designed to have its own characteristics that lead to different user behavior towards app usage. Table V shows the mean value of the number of applications incurred by users. iOS users tend to use most extensive applications. Furthermore, we investigate the user distribution on app diversity. First, we calculate the number of apps incurred by each user, and then we normalize it by the maximum value among the same OS users. If the normalized value of the number of apps incurred by a user is more than 0.5, this user is considered to incur high app diversity. If the normalized value is less than 0.1, this user is considered to incur low app diversity. If the normalized value is between 0.1 and 0.5, the user is considered to incur medium app diversity. The user ratio of each app diversity group in each OS is listed in Table V. Fig. 8 shows the user distribution of app diversity in Table V. The majority of iOS and Windows Phone users incur medium app diversity while the majority of Android users incur low app diversity. This observation indicates that a few Android users have used a considerably large number of applications as compared to the majority of Android users. However, the majority of the other two OS users have used the similar number of applications.

TABLE V
USER RATIO OF EACH APP DIVERSITY GROUP

OS	App Num	HD	MD	LD
Android	5.47	0.004%	33.73%	66.27%
iOS	7.80	1.24%	87.55%	11.21%
Windows Phone	4.97	2.42%	79.85%	17.73%

VI. Conclusions

In this study, we have presented the traffic and user behavior characteristics of three major smartphone operating systems in cellular networks. All of the three operating systems incur considerably high traffic volume at night and Windows Phone exhibits the biggest traffic variance. In terms of application categories, browser and IM attract more users, thus achieving larger user ratios. The obviously high user ratio of Apple Market indicates Apple Market is popular among users. iOS and Windows Phone users exhibit similar application diversity. Another noteworthy characteristic is that iOS users generate much more traffic volume than the other two OS users.

The OS recognition method of big network data in our study has practical significance in user identification. The traffic analysis can help operators achieve better cellular network optimization. In addition, the application diversity analysis has important implications in developing and recommending applications for specific OS users.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (61072061), 111 Project of China under Grant No. B08004, and the Fundamental Research Funds for the Central Universities (2013RC0114).

References

[1] Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017. White Paper, February 2013.
[2] Apple. Apples App Store Marks Historic 50 Billionth Download. http://www.apple.com/pr/library/2013/05/16Apples-App-Store-Marks-Historic-50-Billionth-Download.html.
[3] J. Dean, S. Ghemawat, MapReduce: Simplified Data Processing on Large Clusters, Google, Inc. October 2004
[4] H. Falaki, D. Lymberopoulos, R. Mahajan, R. Govindan,S. Kandula, and D. Estrin. Diversity in Smartphone Usage. In Proc. ACM MOBISYS, 2010.
[5] G. Maier,, F. Schneider and A. Feldmann, A First Look at Mobile Hand-Held Device Traffic, in Lecture Notes in Computer Science.vol. 6032, pp. 161-170, 2010.
[6] M. Z. Shafiq, L. Ji, A. X. Liu, Jia Wang Characterizing and modeling internet traffic dynamics of cellular devices. Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems. ACM, 2011.

[7] A. Botta, D. Emma, A. Pescape, G. Ventre, Systematic Performance Modeling and Characterization of Heterogeneous IP Networks, Journal of Computer and System Sciences, vol. 72, no.7, pp.1134-1143, 2006.

[8] R. Keralapura, A. Nucci, Z.-L. Zhang, and L. Gao. Profiling users in a 3g network using hourglass co-clustering. Proceedings of the sixteenth annual international conference on Mobile computing and networking. ACM, pp. 341-352, 2010.

[9] Q. Xu, J. Erman, A. Gerber, Z. Mao, J. Pang, and S. Venkataraman, Identifying diverse usage behaviors of smartphone apps, Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. ACM, pp. 329-344, 2011.

[10] Composite Capability/Preference Profiles (CC/PP): Structure and Vocabularies 1.0, W3C Recommendation, Jan. 2004.

[11] OMA User Agent Profile V2.0, Specification of Open Mobile Alliance, June 25, 2007.

[12] R. Fielding, J. Gettys, etc., Hypertext Transfer Protocol – HTTP/1.1, RFC2616, June 1999.

[13] Wireless Universal Resource FiLe open source project (WURFL), http://wurfl.sourceforge.net/.

[14] J. Liu, Y. Li, F.Cuadrado, S. Uhlig and Z. Lei, Parallelized Jaccard-Based Learning Method and MapReduce Implementation for Mobile Devices Recognition from Massive Network Data, China Communications, vol. 10, no. 7, pp.71-84, 2013.

[15] General Universal Mobile Telecommunications System (UMTS) architecture, 3GPP Technical Specification 23.101, Dec. 2004.

[16] C. Wang and X. Zhou, "Design of P2P Traffic Identification Based on DPI and DFI," International Symposium on CNMT 2009, Jan.2009, pp.1-4.

[17] IMEI Allocation and Approval Guidelines, GSM Association Official Document, July 27, 2011.