Subscriber Profiling for Connection Service Providers by Considering Individuals and Different Timeframes

Article *in* IEICE Transactions on Communications · June 2016

DOI: 10.1587/transcom.2015EBP3467

Kasim Oztoprak
Konya Food and Agriculture University

IEICE TRANSACTIONS

on Communications

VOL. E99-B NO. 6
JUNE 2016

A PUBLICATION OF THE COMMUNICATIONS SOCIETY

The Institute of Electronics, Information and Communication Engineers
Kikai-Shinko-Kaikan Bldg., 5-8, Shibakoen 3chome, Minato-ku, TOKYO, 105-0011 JAPAN

PAPER

Subscriber Profiling for Connection Service Providers by Considering Individuals and Different Timeframes

Kasim OZTOPRAK[†a)], ***Member***

SUMMARY Connection Service Providers (CSP) are wishing to increase their Return on Investment (ROI) by utilizing the data assets generated by tracking subscriber behaviors. This results in the ability to apply personalized policies, monitor and control the service traffic to subscribers and gain more revenue through the usage of subscriber data with ad networks. In this paper, a system is proposed to monitor and analyze the Internet access of the subscribers of a regional SP in order to classify the subscribers into interest categories from the Interactive Advertising Bureau (IAB) categories. The study employs the categorization engine to build category vectors for all individuals using Internet services through the subscription. The proposal makes it easy to detect changes in the interests of individuals/subscribers over time.

key words: *subscriber categorization, internet usage characteristics, personalized policies*

1. Introduction

The popularity of the Internet among users increases exponentially because of the unlimited access rights to enormous resources contained. The invention of search engines has allowed the users to access knowledge and information easily and the invention of social media changed the lifestyle of people using the Internet. Online gaming fills the leisure time of Internet users. The use of the Internet has helped us to program all our lives by using our wired or mobile devices. According to Portio Research Report [1], the number of mobile subscribers is almost 92% of the world‘s population in 2013. This trend has attracted CSPs to offer flexible services to conform with the different needs of subscribers employing huge data usage plans.

Due to the dramatic increase in traffic in data networks, CSPs are adopting several systems for controlling and monitoring the data traffic. For this, Deep Packet Inspection (DPI) has become one of the key network intelligence technologies for managing the data traffic. [2]. Controlling the traffic of the subscribers enables (i) applying personalized policies; (ii) monitoring and controlling the service traffic; and (iii) gaining more revenue by using the subscriber data. Although the first two items are related to the service provided to the subscribers, the latter is for the sake of the operators by using the Internet access characteristics of the subscribers. As compared with the subscriber analysis systems through http access; DPI systems can bring additional inline information from all applications in use.

The studies concentrating on Internet usage behaviors and part of policy application systems require Internet category databases either to apply several policies on the category of the destination to access or to put a subscriber into an interest category to deliver related advertisements. Digital marketing is defined as “The use of digital technologies to create an integrated, targeted and measurable communication which helps to acquire and retain customers while building deeper relationships with them” [3]. The SPs (e.g. Google, Facebook, Twitter etc.) are able to profile the subscribers by using their history of Internet usage or the behaviors of them through the Internet content access. Profiling allows us to cluster the subscribers into categories according to their interests obtained from their Internet usage and behaviors. This categorization helps the SPs/CSPs for service personalization, predictive offer management, loyalty management, as well as targeted advertisements.

The click ratio of the users increase when they are exposed to advertisements which are more related to their interests [9]. The amount of this increase can go as high as 60% for the first best and 21% for the second best selection [18]. In the same study, the click rate for random ads is reported to be 26%. It is clear that once the subscribers were classified for specific categories, the click rate for the same subscriber group increased tremendously compared with random advertisement views according to the consumer behavior.

In a previous study by the author of this paper [17], a system was proposed and developed to monitor and analyze the Internet access (http) logs of the subscribers of a regional CSP in order to categorize the subscribers into an interest category (or into several interest categories) from the category database which was built to serve as a helper to the categorization engine. In the design of the proposed system, an Internet category database having more than 150M URLs was used. It was compatible with the IAB [4] category database. In the study, an Internet access vector was formed for each subscriber to count and record the accesses to different categories. Then, all access logs were evaluated using the category database in order to identify the accessed categories. In the final analysis, selecting the most accessed three categories formed the Interest vector for each subscriber. The aim of the study was to find differentiating properties between the subscribers in order to use with advertisement systems. More than 80% of the information was common to all users due to popular content such as search engines, Facebook, etc. In order to differentiate between the

Manuscript received November 3, 2015.
Manuscript revised February 1, 2016.

[†]The author is with the Department of Computer Engineering, KTO Karatay University, Konya, Turkey.
a) E-mail: kasim.oztoprak@karatay.edu.tr

DOI: 10.1587/transcom.2015EBP3467

interest categories of the subscribers more accurately, the common traffic was eliminated. In [17], more categories were used for differentiation compared to the existing literature, leading to maximal compatibility with IAB categories.

It is obvious that the interests of any subscriber exhibit temporal variability. While a subscriber is interested in cooking meals between 16.00 and 20.00, the interest would concentrate more on online shopping after 20.00. Hourly and daily variability of interests were considered. The different users in a subscriber environment also affect the global interest category of a subscriber. Interestingly, from the results of the previous study [17], the interest concentration into multiple categories. One of the reasons for that was the possibility of having multiple users at any Internet link.

This study is an extension of the previous [17] one and gives more details about the subscribers by considering the time spent on the pages and the online applications used such as gaming and torrents, which do not reflect real interest in http logs. The respective data for that purpose was taken from the DPI systems indicating the time spent and amount of traffic used for such applications. This data was taken so that more precise information on the subscribers could be obtained. Secondly, the timeframes for interest change were considered and the interests for a subscriber were classified into several categories: all time, last month, last week, last 24 hours, weekend and prime time interests. This was achieved by employing a sliding window method in Internet usage appropriate to the interest groups. In addition, by utilizing the data from the DPI systems, the applications used by subscribers were also considered in order to calculate the interest categories of the subscribers. Finally, a method was employed to distinguish different users from an Internet line and to treat them as different subscribers to perform a better classification.

The rest of the paper is organized as follows. In Sect. 2, a brief summary of the related literature is summarized. Background definitions and a formal description of the proposed solution to the categorization problem is defined in Sect. 3. Experimental results and the comments of the author on the topic are presented in Sect. 4. Finally, the conclusions and future approaches of the topic are given in Sect. 5.

2. Internet Advertisement Systems

In the literature there are several studies concentrating on increasing the revenue of SPs and CSPs. The main target of the revenue increase studies is profiling subscribers. Marketing papers in the past were written to report the research results aiming to investigate alternative methods and complex models of estimating consumer demands [7]. The author offered a method to cluster subscribers into four separate market segments based on the degree of consumer interest for a new product by utilizing canonical correlation.

While information technologies are becoming more and more pervasive, studies have started to collect basic information from our shopping habits. Johansen et al performed a study investigating which foods are bought together [8]. The study was limited to a specific domain and investigated whether people who buy wine buy healthier food items than those who buy beer. The study gives an idea about the trends in marketing science.

The researchers from the marketing and computing domains started to cooperate to find solutions to the profit increase problem. The authors in [3] made a study of the literature delineating the details of the problem from the expectations of marketplace and the ability of big data analytics. They defined the marketing objective as designing a marketing mix that precisely matches the expectations of customers in the targeted segments. The segmentation was to classify the consumers into different groups according to their interests. Interestingly, they pointed out the need of having professional people who have skills to understand the dynamics of the market and can identify what is relevant and meaningful.

The use of the Internet and social media increases interest of the manufacturers to get the consent of the customers by interacting with them and mining the data collected through social media. Turban et al [14] examined the social media commerce and marketing performed by Sony after facing a considerable decline between 2008 and 2012. Portals, customer relationship pages and blogs, twitter and discussion groups were monitored and the data extracted through them were mined to get the consent of the subscribers and keep interacting with them. The results showed that interacting with customers increased the click ratio by 22%. Page views, conversation rates and engagement activities increased the click ratio by 100%. The recent literature review on Internet marketing [15] presented that purchase intention and social media are extremely related to each others. In the literature [16], consumer-purchasing process is classified into several steps: (i) Need recognition, (ii) search for product information, (iii) product evaluation, (iv) product choice and purchase, and (v) post purchase use. The aim of the marketing strategists is to enter the chain from the appropriate point, by advertising the product to the correct people at the right time. Although knowing where to attack is useful, it does not matter much if you cannot decide whom to attack. The marketing strategies are determined by sociocultural environment and one‘s spending capacity.

With the increase in the popularity of the Internet, the expectations and strategies of the marketing domain started trying to find a position in Internet environment through advertisements. Banner ads can be accepted as a primitive form of web advertisements. They were popular between 1995–2001 [9]. The websites were charging advertisers for every page view. The main problem of the banner ads was the untargeted viewers generating a limited amount of clicks without the ability to return the investment. To compensate for the problems incurred by untargeted banner ads, performance based advertisement systems were developed. In such systems, advertisers are responsible for paying only when their advertisements are clicked. In order to increase the revenue from the advertisement system, it targets the subscribers/users having more interest in the advertisement content. Then, the problem becomes one of classification of subscribers into their interest classes. In [18], the authors performed a study on the alternative ad placement regimes. The results indicate that the click through rate for a subscriber can increase up to 60% if information is available to target the placement to a specific consumer.

Literature offers several studies concentrated on mass volume data analysis using big data systems to classify Internet users (consumers) into several categories. The aim of the classification is to increase revenue from the subscribers by advertising the products related to their interests. The advertisement delivery systems work on different criteria like Internet usage, locations, etc. Hall and Kanar [10] have a patent for their system to deliver advertisements according to the locations of the subscribers through mobile telephony networks. They keep an interest database of the mobile subscribers; considering the location information, the advertisements are delivered to the subscribers. As in [11], there are also lots of studies showing an affiliate network how to deliver advertisement data to the customers.

Reference [12] proposed an end to end solution to gather subscriber data from the mobile operator network; they analyzed them through big data systems and then they classified the subscribers into several categories, and finally they delivered advertisements according to the interest categories of customers. According to the model presented, the data were collected through telecommunication system subscribers including broadband, mobile, and IPTV subscribers. They aggregated the data into several big data systems and then classified the subscribers into their respective categories. Finally, a mechanism to deliver advertisement data from the marketers to the subscribers was offered by the authors. Although the design and the study seems perfect, the arguments discussed and the details of the system lack explanations. The design of the study appears to be of a high quality Internet-based advertisement system which combines subscriber interests and advertisement networks.

Since access to web access logs is easy for system administrators, most of the revenue generation systems rely on http/https logs. On the other hand, there is a wealth to be discovered about the subscribers through DPI systems having the application awareness capability. These systems enable the gathering of information for most of the applications used by the subscribers/users. As an example, consider a specific user spending most of his time on a specific game. In spite of this fact, the ratio of his Internet access reporting game access can be less than 5%. Therefore, none of the revenue generation systems classify the subscriber in a gaming category, although s/he should be.

The nature of the Internet traffic pattern was analyzed by Zipf's law [13]. The main idea covered by the law can be adapted to our study in order to confirm the naturalness of the traffic patterns.

3. Subscriber Categorization Engine — SCE

The proposed classification system design, which considers both the Internet access through http systems and the time spent with other applications, is explained in this chapter.

Although traffic classification is a very challenging task in computer networks, the demand to have accurate information on the subscriber's demands and interests in the Internet has triggered the development of several systems for subscriber classification. The early tools classified the subscribers according to their gender, age, marital status, location, etc. which is already available in Customer Repository Management (CRM) systems. The information extracted from the CRM systems does not fully satisfy the expectations of the advertisement networks. In this study, the SCE has been modeled and developed as an approach to get more useful data on the subscribers in order to direct them to appropriate advertisements and to gain useful intelligence about the subscriber that is under consideration.

The SCE consists of two main components. First is the category database with classifications of the URLs having definitions of 121 categories. The second component is the categorization engine aggregating category database and usage logs together to build the interest vectors of the subscribers.

3.1 URL Category Database

Internet category databases have recently been used in many different areas: (i) to provide a knowledge repository for policy enforcement systems for subscribers to allow/deny accessing to URLs and (ii) to build interest vectors/matrices for subscribers to be used by affiliate networks.

The study of building an Internet Category Engine (ICE) is not limited to this study. The work started by developing a regional language and a culturally-intensive Internet category database. The accuracy and granularity of the data sets stored in the ICE were designed to provide better contextual insight of web pages and content; more effective targeting by matching web content to user profiles; and real-time web filtering decision support for policy enforcement systems.

Although [4] offers 26 Tier 1 (top categories) to conform with the IAB standards; ICE has 37 Tier 1 categories with more granular top domain categories. While there are also 364 Tier 2 categories in the IAB database, there are 121 categories in the ICE. However, after excessive experimentation, it was seen that the gap between the two category databases does not affect the advertisers, since most of the advertisers concentrate on several subcategories which are all covered by the ICE, as well as by the IAB category database.

Technology, Software and Services, Business and Financial Services are three of the 37 main Tier 1 categories. Technology and Computer, Search Engines, Content Delivery Networks (CDN) are examples of Tier 2 categories. This is done by adding/updating and/or deleting the 150K+ domains in order to keep the domain listing and categories up to date for the current 150M+ domains in the ICE. In contrast to reports in [5], the number of active domains was reported to be less than 260 million. In addition, the results

Table 1 Different time-zones for experiments.

Zone Name	Timeperiod
WZ0	Weekend: Between 10.00 and 20.00
WZ1	Weekend: Between 20.00 and 10.00
Z1	Weekdays: Between 08.00 and 22.00
Z2	Weekdays: Between 22.00 and 08.00
Z3	Last Week (Cumulative)
Z4	Last Month (Cumulative)
Z5	All Records in the System (Cumulative)

obtained from different experiments showed that 95% of Internet traffic was generated from the top 1 million domains. The ICE database was stored in MySQL to keep the domain lists to help the mapping into categories, and to identify the relations among the categories. Elastic Search was used to query from the database, which allows us to keep a specified amount of URLs in memory, and form indexes in memory as well. The final results are kept in a reporting server where NoSQL runs on them to serve as a very fast medium for reporting queries. The ICE engine was capable of re-querying a URL from the Internet for uncategorized URLs to classify them into an appropriate category.

3.2 Interest Matrix and Interest Vector Generation

One of the most serious problems of the CSPs is the number of users in an Internet access subscription. Although it seems difficult to detect the number of individuals in an Internet access network, the variety of the operating systems in use helps us to count the number of devices accessing the Internet and differentiate them as individual subscribers. The problem arises when two or more devices use the same network. The solution to this problem is concentrating the sequence numbers in TCP sessions. Since, they start randomly, it is too difficult to overlap for two devices having very close sequence numbers with the same version of the operating system. By tracking the operating systems with the TCP sequence number, we can identify the individuals behind the subscriptions. The second improvement in SCE was to give different interest categories for subscribers according to different timeframes. The system is designed to handle different timeframes depicted in Table 1.

The second part of the study demonstrates how to form interest-frequency vectors for each subscriber in order to generate interest matrices. During this study, a person is identified in all subcategories. The interest matrix is a table holding the access count of every subscriber in each category. The matrix has a dimension of MxN where M is the number of subscribers while N is the number of categories. Interest matrix entries for each subscriber are processed to build the interest ratio vector containing the three most popular categories of subscribers. As can be seen Fig. 1, the interest matrix generation algorithm works as follows:

- Form 122 columns in an intermediate table to count the total accesses for each category and the sum of the total access for every subscriber,
- Form a popular interest ratio vector holding the three most popular interest category access ratios for every subscriber,
- Collect the logs for the Internet access of subscribers,
- Parse the logs and get the top level domains and subdomains if applicable,
- Select the subscriber from the parsed data and increase the count for the fields representing the total access and appropriate interest column in the interest vectors after querying the domain from the ICE to match,
- Send a request to the Internet to get the page accessed, if the domain is not listed in the ICE. After getting the content of the accessed URL, the page is classified according to Tier 2 categories and inserted into the ICE database,
- Hash the top 10 domains, and keep them in a cache in order to accelerate the categorization. The query is first forwarded to the hashed top 100 domains; if it does not succeed, then a query is performed through the ICE,
- Update the interest ratio vector by another thread to keep the categories for a subscriber up to date.

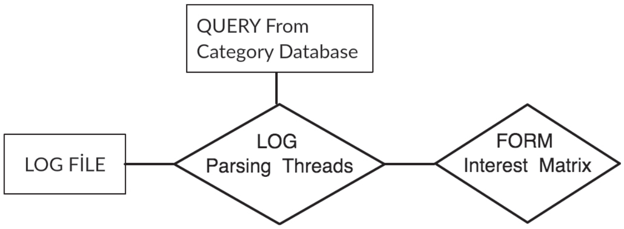

Fig. 1 Basic architecture of interest matrix generation algorithm.

Table 2 Comparison of the literature, Ref. [17] and the SCE.

	Current	Ref. [17]	The SCE
Demographic Data	Yes	Yes	Yes
Interest Classification	No	Yes	Yes
Interest in Different Times	No	No	Yes
Individual Identification	No	No	Yes
Application Aware Cat.	No	No	Yes

Although the algorithm above seems relatively easy, NoSQL was used to conform to the expected processing power in the reporting engine. Table 2 summarizes the contribution of this study by comparing the work done in the previous study [17] and the currently available systems.

4. Experimental Results

The experimental studies and numerical experiments conducted for the designed system are presented in this part. Since building of the ICE database is not the main scope of this study, there are no experiments conducted on building of the ICE database. The ICE database runs on an Intel-based system having two 8-core Xeon Processors with the speed of 2.4 GHz and 256 GB of main memory. There are also 2 TB of SAS disk and 128 GB of SSD disk space. The categorization engine runs on a single server with the same configuration of the ICE.

The experiments were performed using two different approaches. The first scenario involved the intensive Internet access of university students through mobile devices where there was no need for detection for usage of multiple devices. There was also no need for the category evaluation for different timeframes since the usage was recorded during the daytime from 09.00 to 18.00. The next scenario involved the temporal change of the Internet usage characteristics and the detection of the number of different users through a single subscription.

It is obvious that building such a system requires dynamism and responsiveness in updating the Internet category database and delivering the response to the requests by the parsers of the log processors. The final part of the experiments contains the information on the database update and the scalability of the system.

4.1 Single User through a Subscription (an Access)

The results reported in this subsection are largely adapted from [17]. The experiments in this step were performed on the Internet access of university students over a period of 75 days. Most of the users used their smart phones to connect to the Internet thus affecting the characteristics of the traffic patterns. The experimental results were classified into two main groups. The first concentrated on system-specific results indicating the amount of total traffic, the most popular domains accessed by the system subscribers, the total web pages, and uncategorized domains that were accessed. The second group focused on subscriber-level details, which presented user-interest categories.

The total number of subscribers involved in the experiments was 2,833. The number of total page visits during the experiments was 2,101,850. The number of connections for different objects was 32,777,581. The average number of web pages visited per month for a subscriber was 297. However, there were 15.59 URLs on the log for a unique web visit processing a monthly total of almost 4,630 entries per subscriber monthly. The world average for a unique web visit was 2,278 per month in 2013 [6]. The total traffic consumed by the subscribers during the experiments was 2,589.92 GB. In the experiments, although the number of accessed domains was 44,898, it was realized that 61.7% of the requests included the top 10 domains. The ratio for top 100 domains was 88.3% and 95% of the traffic was to the first 1,000 top domains. In order to investigate the behavior of the subscribers with different interest categories, the common domains and interests like search engines, similar news pages, etc. were excluded. At the end of the categorization, interest vectors for 2,833 subscribers were created. In the categorization process, it was realized that 93% of the domains were directly resolvable through the ICE database. The dynamic categorization feature was not activated during the analysis of the logs mentioned hence the unknown domains.

Table 3 presents the results for the most actively accessed Internet categories according to the number of accesses and the ratios of the total accesses. The total traffic consumed to access those domains is presented in Table 4. Naturally, the most popular domains according to the number of connections were not same as the ones reported by the consumed data. The results show that the majority of users clustered around the categories of Technology and Computer, Online Video and Search Engines. Accessing the pages classified into those three categories consumed more than 50% of the traffic. In addition to presenting the graph representation of Table 3 and Table 4, Fig. 2 and Fig. 3 also compare the Zipf's distribution within a larger data set.

Table 3 Top 5 interest categories according to number of connections.

Categories	# of connections	Connection ratio
Image Sharing	7,864,377	23.99%
Tech & Comp	3,454,700	10.54%
Advertisements	2,698,199	8.23%
Unknown	2,513,749	7.67%
News	2,449,113	7.47%

Table 4 Top 5 interest categories according to data transferred (in gigabytes).

Categories	Data Transferred	Transfer Percentage
Tech & Comp	553.50	21.37%
Online Video/Audio	489.079	18.88%
Search Engines	333.236	12.86%
Internet Portals	227.977	8.80%
Image Sharing	195.09	7.50%

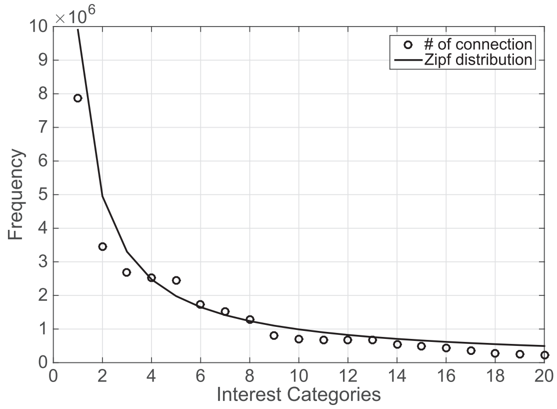

Fig. 2 Top 20 interest categories according to number of connections (in thousands) and related Zipf distribution.

Tables 5–7 present the results for the Internet categories in which subscribers were assigned according to their Internet usage characteristics. For a subscriber, the number of accesses within a category is divided by his/her total number of accesses. The results for each category group was calculated; after the calculations, the interest ratio vector was sorted in descending order to present the most popular interest category for that user.

Table 5 presents the number of subscriber assignments according to their first preferences. The experiments for this category were repeated twice with the threshold values 0.5 and 0.35 respectively. Almost half of the subscribers

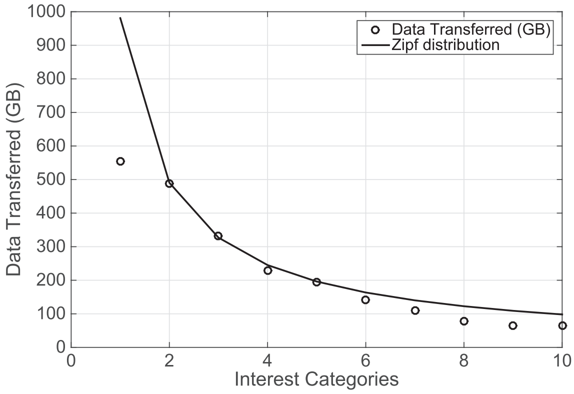

Fig. 3 Top 10 interest categories according to the data transferred (in gigabytes) and related Zipf distribution.

Table 5 Top 5 interest distributions according to the first preferences of subscribers.

Categories	*Ratio* > 0.5	*Ratio* > 0.35	Categories
Image Sharing	438	756	Image Sharing
Tech & Comp	98	229	Tech & Comp
Search Engines	44	109	Search Engines
Unknown	25	53	News
Games	23	51	Unknown

Table 6 Top 5 interest distributions according to the second preferences of subscribers.

Categories	*Ratio* > 0.3	*Ratio* > 0.2	Categories
Tech & Comp	31	163	Tech & Comp
Search Engines	17	105	Search Engines
Image Sharing	17	98	Image Sharing
Unknown	14	69	Unknown
CDN	5	47	News

Table 7 Top 5 interest distributions according to the third preferences of subscribers.

Categories	*Ratio* > 0.2	*Ratio* > 0.1	Categories
CDN	8	282	Advertisements
Tech & Comp	5	225	Search Engines
Image Sharing	4	192	Tech & Comp
Unknown	4	150	Unknown
Search Engines	4	97	News

had a choice with an interest ratio higher than 0.35. Interestingly, when the frequency ratio changed, the popularity order of the interests also changed. The same experiments were repeated for the second and third most popular interest categories and the results are summarized in Tables 6 and 7, respectively. The experiments were repeated with changing frequency ratios for both experiments.

Figure 4 presents the data in Tables 5–7 in a graphic form. While second and third preferences have similar behaviors that start from small values and decrease gradually, the first interest line performs an attitude similar to Zipf's function.

It is important to differentiate the subscribers to attract the advertisers. Concentrating on rarely accessed websites and blogs we can only understand that these subscribers are different. This reality has been observed during the calculations performed. Top 25 domains such as google.com, mail.ru, apple.com, and instagram.com were excluded from the interest classifications in order to differentiate subscribers. In addition to the famous sites, search engines and uncategorized domains were also excluded from the database.

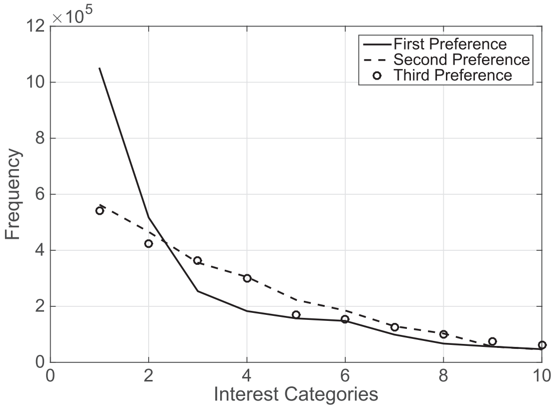

Fig. 4 Top 10 interest category subscriber frequencies according to the first/second/third preferences.

Table 8 Top 5 interest categories according to the number of connections after purification.

Interest Categories	# of connections	Connection Ratio
Technology and Computer	2,091,681	15.58%
News	1,897,084	14.13%
Shopping	874,603	6.51%
CDN	803,592	5.99%
Business Services	653,905	4.87%

Table 9 Top 5 interest categories according to data transferred (in gigabytes) after purification.

Interest Categories	Data Transferred	Connection Ratio
Tech & Comp	98.73	15.71%
Online Video/Audio	92.88	14.78%
CDN	72.21	11.49%
Pornography	58.16	9.25%
News	46.69	7.43%

After performing the elimination of access data to top 25 domains, the number of subscribers decreased to 2,811 in the remaining access file. The number of total page visits during the experiments decreased to 828,448; however, the number of connections for different objects decreased to only 13,426,026. Although, the decrease in most of the parameters excluding the number of subscribers was about 60%, the total traffic consumed by the subscribers aggressively decreased to 628.39 GB. Naturally, the number of unique domains accessed decreased to 31,811. The second interesting result was observed when the traffic of popular domains was investigated. The traffic consumption ratio for the top 10, 100, and 1000 domains decreased to 22.98%, 71.42%, and 95.03% respectively.

Tables 8 and 9 are purified counterparts of Tables 3 and 4 respectively. One of the most important findings in the

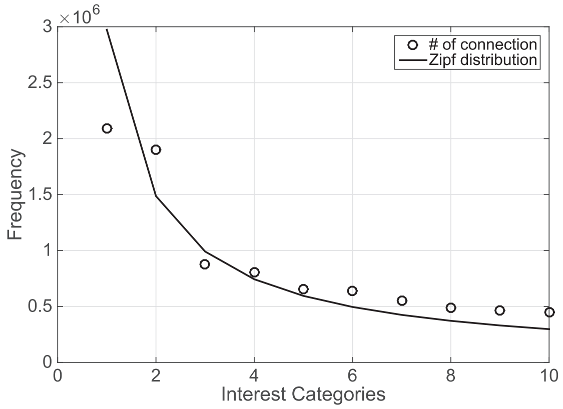

Fig. 5 Top 10 interest categories according to number of connections (in thousands) and related Zipf distribution after purification.

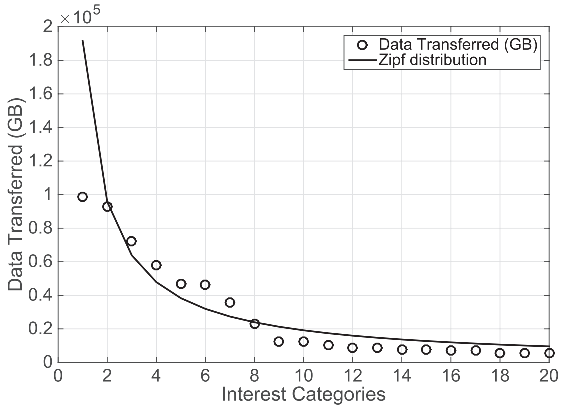

Fig. 6 Top 20 interest categories according to the data transferred (in gigabytes) and related Zipf distribution after purification.

Table 10 Top 5 interest distributions according to the first preferences of subscribers after purification.

Categories	$Ratio < 0.2$	$Ratio > 0.35$	**Categories**
Tech & Comp	336	86	Tech & Comp
News	152	35	News
Mobile	89	31	Games
Shopping	63	13	CDN
CDN	58	11	Shopping

experiments was having a more divergent distribution on the interest categories according to the number of connections and the traffic incurred for the pages. Furthermore, differentiating characteristics were observed by including different interest categories like Education, Marketing, and Pornography. Although there is diversity in the interest category distribution, the system naturally adapts itself to Zipf's distribution after excluding a few initial samples as can be seen in Figs. 5 and 6.

Tables 10–12 demonstrate the results of interest category assignment vectors after the purification of the common traffic. In contrast, in Tables 5–7, the results present the diversity of the values obtained from the experiments which are the preferred behaviors to run with an advertisement network. Figure 7 also confirms the results of Tables 10–12 when compared with Fig. 4 as well.

Table 11 Top 5 interest distributions according to the second preferences of subscribers after purification.

Categories	$Ratio < 0.1$	$Ratio > 0.2$	**Categories**
Tech & Comp	548	15	Tech & Comp
News	216	10	News
Mobile	144	5	CDN
CDN	136	4	Government and Org.
Shopping	123	3	Mobile

Table 12 Top 5 interest distributions according to the third preferences of subscribers after purification.

Categories	$Ratio < 0.1$	$Ratio > 0.1$	**Categories**
Tech & Comp	548	15	Tech & Comp
News	216	10	Shopping
Mobile	144	5	News
CDN	136	4	Mobile
Shopping	123	3	Business Services

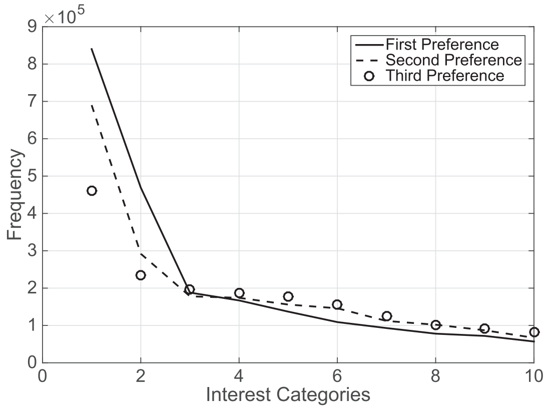

Fig. 7 Top 10 interest category subscriber frequencies according to the first/second/third preferences after purification.

4.2 Multiple Subscriber Detection and Changing Interest Distribution

The experiments in this step were performed on the Internet accesses from a CSP for 60 days. Although a wired connection was used, most of the user devices accessed the Internet via mobile devices. The experiments provided information on both systemwise and subscriberwise statistics.

The experiments were applied to a small network node; the total number of subscribers involved in the experiments was 1,117. The number of total page visits during the experiments was 7,342,389. The average web page visits per month for a subscriber was 3,287. The total traffic consumed by the subscribers was 107.548 TB. In the experiments, although the number of accessed domains was 69,887, it was realized that, 63.2% of the requests were to the top 10 domains. The ratio for the top 100 domains was 86.3% and 94.7% of the traffic was to the first 1,000 top domains. The number of different devices used to access the Internet through the experimented subscriber lines was 4,325 giving an average of 3.87 devices per line. The number of mobile devices used

Table 13 Top 5 interest categories according to number of page visits.

Interest Categories	# of connections	Connection Ratio
Technology and Computer	277,542	21.0%
News	158,595	12.0%
Shopping	100,443	7.6%
Social Networks	85,774	6.49%
CDN	68,724	5.2%

Table 14 Top 5 interest distributions according to all time and last month preferences of subscribers.

Categories	All Time	Last Month	Categories
Tech & Comp	479	501	Tech & Comp
News	291	317	News
Mobile	212	211	Social Networks
Shopping	177	162	CDN
CDN	148	93	Games

Table 15 Top 5 interest distributions according to last week and weekend preferences of subscribers.

Categories	Last Week	Weekend	Categories
Tech & Comp	496	397	Tech & Comp
News	315	342	News
Social Networks	198	276	Social Networks
CDN	192	184	Games
Shopping	134	153	Sports

Table 16 Top 5 interest distributions according to different timeframes on weekends.

Categories	Daytime	Nights	Categories
Social Networks	295	312	Social Networks
Mobile	191	278	Shopping
Shopping	286	257	Games
Tech & Comp	201	243	Tech & CDN
News	188	187	Sports

was 2,835. Interestingly, the number of common domains for Internet access was 49. The same purification steps were applied as in the first part of the experiment.

After performing the purification steps, the number of subscribers decreased to 4,234. The number of total page visits during the experiments decreased to 1,321,630. Although the decrease in most of the parameters excluding the number of subscribers was about 82%, the total traffic consumed by the subscribers aggressively decreased to 7.53 TB meaning that 93% of the traffic was purified. Naturally, the number of unique domains accessed decreased to 57,832. A second interesting result was observed when the traffic of popular domains was investigated. The traffic consumption ratio for the top 10, 100, and 1000 domains decreased to 26.44%, 69.37%, and 93.81% respectively.

Tables 14–17 show the results of interest category assignment vectors after the purification of the common traffic. The results present the changing nature of the interest assignment for changing timeframes. For example, weekends and night usage patterns are more leisure oriented when compared with the other tables. It is clear from Table 16 that the users show more divergent behavior on weekends.

Table 17 Top 5 interest distributions according to daytime and night time preferences of subscribers.

Categories	Daytime	Night	Categories
Tech & Comp	611	447	News
News	364	401	Social Networks
Social Networks	227	389	Tech & Comp
CDN	179	166	CDN
Education	99	142	Image Sharing

During the experiments the effect of DPI systems were also considered. If an individual used an application (games, online video etc) more than 25% of the total Internet usage time, his/her interest was set to that category. The results of both experiments indicated that the efforts performed during the studies gave promising results since the system successfully classifies the subscribers and the individuals using the subscription lines.

Although the system categorizes the individuals behind a subscription line successfully, the problem arises when sending advertisements to them. It will be harder to differentiate between individuals when they are using identical devices and the same operating systems. If the device type or the operating system version were different, inspecting http access patterns can easily solve the problem. Otherwise, the solution to the problem does not exist yet without keeping the medium access control address of the devices, which is not possible with the existing Internet technology. The second problem with individual user detection arises when an individual uses more than one device to access the Internet. The system will be capable of combining the access from different devices through a subscription line by analyzing the Internet access patterns and category assignments. Finding a solution to the latter is easier than to solve the problem of the former.

The calculation of the interest helps the marketing teams of the CSPs to select appropriate advertisements in order to attract the subscribers to click the advertisement. Knowing the tendencies and buying habits of the subscriber‘s increases the revenue generated from the ad networks. The capability of the system increased when compared with the study in [17]. The system has the capability to attract the users with more precise advertisements according to detailed classifications.

4.3 Scalability of the System

In order to evaluate the scalability of the system, a synthetic log file is generated by repeating the Internet access entries from the original access file by one million times. The synthetically-generated log file simulates the Internet access for almost 1.1 million subscribers for 60 days with similar Internet access loads. Similarly, it contains access data for 6.7 million (1.117 * 1M * 60/10) subscribers for 10 days. The categorization and building interest ratio vector processes took a total of 14 days using two servers, and 10 days using three with the features described. The results indicate that, the system with three servers can process the access logs of 6.7 million subscribers in a day. This means that the above system with three servers can process the data of 6.7 million subscribers in almost real time. Since the system uses big data components and the data is suitable to processed in parallel, the system can scale up to tens of millions subscribers.

5. Conclusion and Future Studies

In this study, the effect of categorizing subscribers for ad networks was investigated. Unlike conventional ad networking systems, the subscribers were categorized according to a previously developed system compatible with IAB categories. The aim of the study was to find differentiating points in subscribers to advertise appropriate products in order to make additional revenue from the advertisement systems. The system presented is also capable of defining interest categories for different timeframes, which can easily attract the advertisers. Besides categorization using web access logs with different timeframes, the system is also capable of using the information extracted from the applications such as gaming and social media. Additionally, the experiments conducted throughout the study were performed on the Internet access for all months, during which the changing behavior of the individuals was observed. It is noticed that the traffic pattern for the same individual was different in different timeframes of the day and even during different days in a week. The model proposed in this study provides the capability of computing the optimal advertisement targets for advertisers and CSPs.

One of the most important problems of the system discussed is rooted from the nature of the Internet: the encrypted traffic. There is no unique and direct way to get information from encrypted traffic. Additionally, the SSL/TSL protocol is an obstacle to the usage of information extracted from search keys and social media. In future studies, the possibility of the information retrieval from such systems and the correlation/diversity of mobile subscribers and landline subscribers will be investigated.

References

[1] Portio Research Ltd. Portio Research Mobile Factbook 2013, Report accessed online on Aug. 1, 2015. http://www.portioresearch.com/en/free-mobile-factbook.aspx

[2] Y.J. Lee. J. Oh, J.K. Lee. D. Kang, and B.G. Lee, "The development of deep packet inspection platform and its applications," 3rd International Conference on Intelligent Computational Systems (ICICS'2013) Hong Kong, China, Jan. 2013.

[3] K. Grishikashvili, S. Dibb, and M. Meadows, "Investigation into big data impact on digital marketing," International Conference on Communication, Media, Technology and Design, pp.146–150, 2014.

[4] The Interactive Advertising Bureau (IAB) 2014, API Specifications accessed online on Aug. 1, 2015: http://www.iab.net/media/file/OpenRTB_API_Specification_Version_2_3_1.pdf

[5] Netcraft. The August 2015 Web Surwey, accessed online on Aug. 1, 2015. http://news.netcraft.com/archives/category/web-server-survey/

[6] The Canadian Internet, 2015. Resources accessed online on Aug. 1, 2015: http://cira.ca/factbook/2014/the-canadian-internet.html

[7] K.D. Bahn, "Characterizing consumer interest through the use of canonical correlation: Application for small business," Proc. 1982 Academy of Marketing Science (AMS) Annual Conference, Developments in Marketing Science: Proc. Academy of Marketing Science, pp.519–524, 2015.

[8] D. Johansen, K. Friis, E. Skovenborg and M. Grønbæk, "Food buying habits of people who buy wine or beer: cross sectional study," BMJ, vol.332, no.7540, pp.519–522, March 2006.

[9] J. Lescovec, A. Rajaraman, and J. Ullman, "Mining of Massive Datasets," retrieved from http://www.mmds.org on Aug. 2015.

[10] M.D. Hall and N. Kanar, "Method delivering location-base targeted advertisements to mobile subscribers," US Patent: US7027801 B1, April 2006.

[11] P. Scalise, "Internet affiliate network marketing system and method with associated computer program," US Patent Application: US2015/0142585 A1, May 2015.

[12] J. Wilson, C. Kachappilly, R. Mohan, P. Kapadia, A. Soman, and S. Chaudhury, "Real world applications of machine learning techniques over large mobile subscriber datasets," Feb. 2015.

[13] L.A. Adamic and B.A. Huberman, "Zipf's law and the Internet," Glottometrics 3, pp.143–150, 2002.

[14] E. Turban, D. King, J.K. Lee, T.-P. Liang, and D.C. Turban, "Social commerce: Foundations, social marketing, and advertising," Electronic Commerce, Springer Texts in Business and Economics, pp.309–364, 2015.

[15] T. Natarajan, S. Balasubramaninan, J. Balakrishnan, and J. Manickavasagam, "The state of Internet marketing research (2005–2012): A systematic review using classification and relationship analysis," Marketing and Consumer Behavior: Concepts, Methodologies, Tools, and Applications: Concepts, Methodologies, Tools, and Applications, IGI Global, pp.282–305, 2014.

[16] "Marketing Principles," from http://2012books.lardbucket.org/ on 23rd of Oct. 2015.

[17] K. Oztoprak, "Profiling subscribers according to their Internet usage characteristics and behaviors," Proc. 2015 IEEE International Conference on Big Data (Big Data), pp.1492–1499, 2015.

[18] P. Jeziorski and I.R. Segal, "What makes them click: Empirical analysis of consumer demand for search advertising (July 25, 2012)," accessed online on Jan. 22, 2016 at SSRN: http://ssrn.com/abstract=1417625

Kasim Oztoprak is affiliated with the Computer Engineering Department of KTO Karatay University, Konya, Turkey. He received his BS, MS, and PhD degrees in computer engineering from the Middle East Technical University in 1996, in 2000 and in 2008, respectively. He is a member of The Institute of Electronics, Information and Communication (IEICE) and IEEE. His research interests include computer and communication networks, communication security and performance analysis.