Profiling Subscribers According to Their Internet Usage Characteristics and Behaviors

Conference Paper · October 2015

DOI: 10.1109/BigData.2015.7363912

Kasim Oztoprak
Konya Food and Agriculture University

Profiling Subscribers According to Their Internet Usage Characteristics and Behaviors

Kasım Öztoprak
Department of Computer Engineering
KTO Karatay University, 42020, Konya, Turkey
kasim.oztoprak@karatay.edu.tr

Abstract—Providers (SP) are wishing to increase their Return of Investment (ROI) by utilizing the data assets generated by tracking subscriber behaviors. This results in the ability of applying personalized policies, monitoring and controlling the service traffic to subscribers and gaining more revenues through the usage of subscriber data with ad networks. In this paper, a framework is developed to monitor and analyze the Internet access of the subscribers of a regional SP in order to categorize the subscribers into an interest category from The Interactive Advertising Bureau (IAB) categories. The study employs the categorization engine to build category vectors for all subscribers. The simulation results show that once a subscriber has been classified into a category the click rate for the same subscriber group can be improved by correlating the interests of the subscribers with the advertisements.

Keywords—Subscriber Categorization, Internet Usage Characteristics, Personalized Policies

I. INTRODUCTION

The popularity of Internet among users increases exponentially because of the unlimited access rights to enormous resources contained. The invention of search engines allowed the users to access the knowledge and information easier than it was and the invention of social media changed the lifestyle of people using Internet. Online gaming fills the leisure time of Internet users. The use of Internet helped us to program all our lives by using our wired or mobile devices. According to [1], the number of mobile subscribers is almost 92% of the people living in the world in 2013. This trend attracted Connection Service Providers (SP) - CSP (Telecommunication Operators) to bring flexible services to conform different needs of subscribers employing huge data usage plans.

Due to the dramatic increase of traffic in data networks, service providers are adopting several systems for controlling and having the visibility of the data traffic. By enabling the operators to have the control and visibility on the subscriber data Deep Packet Inspection (DPI) technology has become one of the key network intelligent technologies in order to manage the data traffic [2]. Controlling the traffic of the subscribers brings the ability of i) applying personalized policies; ii) monitoring and controlling the service traffic; iii) gaining more revenues 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 analyze 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 requires 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 service providers (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.

In this study, a framework is developed to monitor and analyze the Internet access 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 an helper to categorization engine. Since most of the advertisement applications are using the category database of The Interactive Advertising Bureau (IAB) [4], the one used in this experiment is almost compatible with the IAB category database. The simulation results show that once a subscriber has been classified into a specific category (or to several categories) the click rate for the same subscriber group increases tremendously compared with random advertisement views. Since the study employs to use Internet access patterns, the results show that more than 80% of the information is common to all users. In the literature, the number of employed categorization was very small. The past researchers do not conduct a full categorization almost identical with IAB categories. Additionally, this specific study is divergent from the others, by employing a mechanism, which only considers uncommon traffic patterns.

The rest of the paper is organized as follows. In section 2, a brief summary of related literature is summarized. Background definitions and formal description of proposed solution to categorization problem is defined in section 3. Finally experimental results and conclusions are given in sections 4 and 5 respectively.

II. Internet Advertisement Systems

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

While the information technologies wraparound us, the studies started to collect basic information from our shopping habits. In [8], the authors performed a study investigating which foods are bought together. 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 marketing domain and computing domain started to cooperate to bring solutions to the profit increase problem. The authors in [3] made a literature survey delineating the details of the problem from the expectations of marketing scope 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 dynamics of market and can identify what is relevant and meaningful.

Banner ads can be accepted as the primitive form of web advertisements. It was popular in time period of 1995-2001 [9]. The websites were charging the advertisers for every impression. The main problem of the banner ads was untargeted viewers generating limited amount of clicks without the ability to return the investment. To compensate the problems incurred by untargeted banner ads, the performance based advertisement systems were developed. In such systems, advertisers are responsible from paying only when their advertisements are clicked. The current systems and search engines are still using same approach. In order to increase the revenue from the advertisement systems, the system targets the subscribers/users with the probability having more interest to the advertisement content. Then, the problem becomes classification of subscribers into their interest classes.

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 revenues from the subscribers by advertising the products related to their interests. The advertisement delivery systems are working on different criteria like Internet usage, locations, etc. [10] has 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 locations information, the advertisements are delivered to the subscribers. As in [11], there are also lots of Internet affiliate network studies offering an affiliate network how to deliver advertisement data to the customers.

[12] come up with an end to end solution to gather subscriber data from mobile operator network; they analyze them through big data systems and then they classify the subscribers into several categories, and finally they deliver advertisements according to the interest categories of customers. According to the offered model, the data are collected through telecommunication system subscribers including broadband, mobile, and IPTV subscribers. They aggregate the data into several big data systems and then classified the subscribers into respective categories. Finally, a mechanism to deliver advertisement data from the marketers to the subscribers is offered by the authors.

Although the design and the study seems perfect, the arguments discussed and the details of the system has lack of explanations and the study seems a high level design study architecting an Internet based advertisement framework by combining subscriber interests and advertisement networks.

Since accessing to web access logs is easy by the system administrators, most of the revenue generation systems rely on the http/https logs. On the other hand, the there is an ocean to be discovered about the subscribers through deep packet inspection (DPI) systems with the ability of application awareness. Those systems bring the ability to get information of almost all applications used by the subscribers/users. A simple example can be given as follows; consider a specific user playing a specific game spending most of his time. In spite of this fact, the ratio of his Internet access reporting game access can be less than 5%. None of the revenue generation systems classify the subscriber in gaming category, although s/he should be.

The use of Internet and Social media increases by the manufacturers to get consent of the customers by interacting with them and mining the data collected through social media. The authors [14] present the social media commerce and marketing performed by Sony after having a considerable decline between 2008 and 2012. Portals, customer relationship pages and blogs, twitter and discussion groups are monitored and the data extracted through them are 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 100%. The recent literature review on Internet marketing [15] presented the concentration of marketing research on Internet marketing. They found out that purchase intention and social media hold high centrality degree among the examined literature.

III. SUBSCRIBER CATEGORIZATION ENGINE ARCHITECTURE- SCEA

Although traffic classification is very challenging task in computer networks, the demand to have accurate information on the subscribers' demands and interests in Internet triggered the development of several systems for subscriber classification. The early tools were classifying the subscribers according to their gender, age, marital status, location, etc. which is already available in the Customer Repository Management (CRM) systems. The information extracted from the CRM systems does not help to the expectations of the advertisement networks. The SCEA has been modeled and developed an approach to get more useful data on the subscribers to direct them to appropriate advertisements and to gain useful intelligence about the “subscriber” that is being studied.

The SCEA consists of two main components. The first component is the category database holding the categories of the domains and separate entries for the domains holding different categories under different URLs. The second component is the categorization engine aggregating category database and usage logs together to build interest vectors of the subscribers.

A. Category Database

Internet category databases are used in many different areas recently concentrating into two: i) to provide a knowledge repository for policy enforcement systems for subscribers to allow/deny accessing to URLs; ii) to build interest vector/matrix for subscribers to be used by affiliate networks.

Our study of building Internet Category Engine (ICE) is not limited to this study. The works started to develop a regional language and culture aware Internet category database. The accuracy and granularity of the data sets stored in the ICE are 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 to have 26 Tier 1 (top categories) to conform IAB QAG 2.0, ICE has 37 Tier 1 categories with more granular top domain categories. While there are also 364 Tier 2 categories in IAB database, there are 121 categories in ICE. However, after deep experimentations, the gap between two category databases does not effect the advertisers, since most of the advertisers concentrate on several subcategories which are all covered by ICE, as well as IAB category database. Although the space does not allow us to mention all Tier 1 and Tier 2 categories, “Technology, Software and Services”, “Business and Financial Services”, “Motors and Vehicles “ are three of top 37 main categories. “Technology and Computer”, “Search Engines”, “Content Delivery Networks (CDN)” are examples to Tier 2 categories. ICE is formed from the scratch and dynamically updated every day by addition/update/deletion of 150K+ domains daily in order to keep the domain listing and categories up to date. On contrary to having about 875M domains reported by the organization [5], currently, there are more than 150M domains in ICE. In contrast to reports in [5], the number of active domains is reported to be less than 260 millions and the number of total domains seems to be 670M. In addition, the experiences obtained from different experiments show that 95% of Internet traffic is to top 1 million domains.

The ICE database is stored in MySQL database to keep the domain lists, the mapping into categories, and the relations among the categories. Elastic Search is used to query from the database, which allows us to keep 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 is capable of re-querying a URL from Internet for uncategorized URLs to classify them into an appropriate category where Latent Dirichlet Allocation (LDA) and similar algorithms are used.

B. Interest Matrix Generation and Update – Categorization Engine

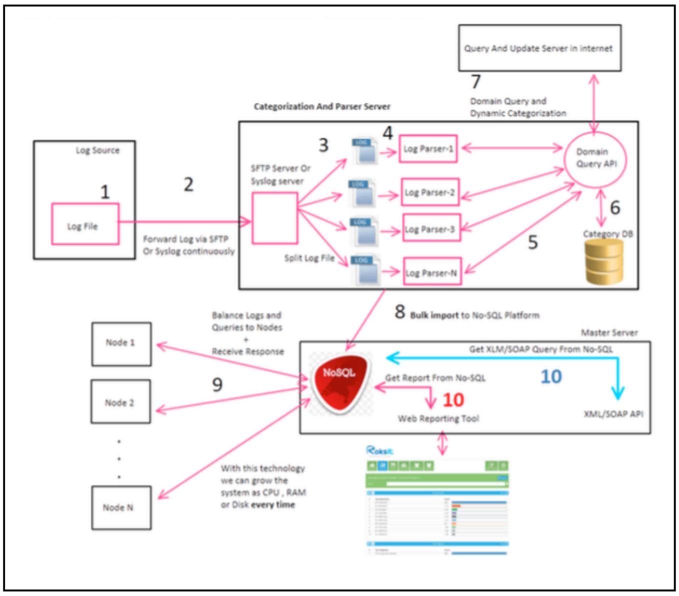


Fig. 1. Basic architecture of interest matrix generation/update

The second part of the study is to form the methods to form interest frequency vectors for every subscriber. During the studies, a person is categorized in all sub categories. As it could be seen from figure 1, the interest matrix generation algorithm works as follows:

- Form additional 122 columns in an intermediate table to count total accesses for every subscriber
- For every subscriber a popular interest ratio vector is formed holding 3 most popular interest category access ratios of the 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 increment the count for the fields representing total access and appropriate interest column in the interest vectors after Querying the domain from ICE to match

- If the domain is not listed in the ICE, a request is sent to Internet to get the page accessed. After getting the content of accessed URL, the page is classified according to Tier 1 and Tier 2 categories and inserted into ICE database
- In order to accelerate the matching subscriber to domain pair, top 100 domains are hashed and kept in caching like environment. The request to match is first forwarded to the hashed top 100 domains, if it does not succeed, then a query is performed through ICE (in the experimental studies it is used as 25 domains, since the domain diversity was limited compared with a Service Provider subscriber domain diversity)
- The interest ratio vector is updated by another thread to keep the categories for a subscriber up to date. This step helps us to keep track of subscriber, to category mappings in the requested form

Although the algorithm above seems pretty easier, NoSQL is used to conform the expected processing power in reporting engine.

IV. Experimental Results

In this part, the experimental studies and numerical experimentations conducted for the designed system is presented. Since building ICE database is not the main scope of this study, there are no experiments conducted on the building of the ICE database. The ICE database is running on an Intel based system having 16 core Xeon Processor with the speed of 2.4 Ghz, 256 GB of main memory, 2 TB of SAS and 128 GB of SSD disks. The categorization engine runs on a single server with the same configuration of ICE.

The experiments are performed on the Internet accesses for a university during 75 days. Most of the users were using their smart phones to connect to the Internet affecting the characteristics of the traffic patterns. The experimental results were classified in two major subjects. The first group concentrated on system specific results indicating the amount of total traffic, the most popular domains accessed by the system subscribers, total web pages accessed, and uncategorized domains. The second part concentrated on subscriber level details presenting user interest categories.

The total number of subscribers reached during the experiments was 2833. The number of total page visits during the experiments was 2101850. The number of connections for different objects was 32777581. Although, the average web page visited per month for a subscriber was 297, there were 15,59 URLs on the log for a unique web visit to be processed accumulating almost 4630 entries per subscriber monthly. The world average for unique web visit was 2278 per month in 2013 [6]. The total traffic consumed by the subscribers was 2589.92 GB. In the experiments, although the number of accessed domains was 44898, it has been realized that, 61.7% of the requests were to top 10 domains. The ratio for top 100 domains was 88.3% and 95% of the traffic was to the first 1000 top domains. In order to investigate the behaviors 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 vector for 2833 subscribers have been created. In the categorization process, it has been realized that 93% of the domains were directly resolvable through ICE database. The dynamic categorization feature was not activated during the analysis of the logs mentioned.

TABLE I. Top 10 interest categories according to the number of connections

Interest Categories	Connection Information	
	# of Connections	***Connection ratio***
Image Sharing	7864377	23.99%
Technology and Computer	3454700	10.54%
Advertisements	2698199	8.23%
Unknown	2513749	7.67%
News	2449113	7.47%
Search Engines	1744826	5.32%
Content Delivery Networks (CDN)	1512204	4.61%
Shopping	1291739	3.94%
Online Video/Audio	795062	2.43%
Social Networks	690436	2.11%

TABLE II. Top 10 interest categories according to the data transferred (in gigabytes)

Interest Categories	Transferred Data Information	
	Data transferred (GB)	***Transfer Percentage***
Technology and Computer	553.50	21.37%
Online Video/Audio	489.08	18.88%
Search Engines	333.24	12.86%
Internet Portals	227.98	8.80%
Image Sharing	195.09	7.50%
Unknown	141.82	5.47%
Content Delivery Networks (CDN)	109.35	4.20%
Social Networks	77.65	3.00%
Games	66.03	2.55%
Online Storage	64.99	2.50%

Table 1 presents the results for most actively accessed Internet categories according to the number of accesses and the ratio of the total access. The total traffic consumed to access to those domains are presented in table 2. 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 major crowd of the users gathered around Technology and Computer, Online Video and Search Engines. Those three categories access consumed more than 50% of the traffic.

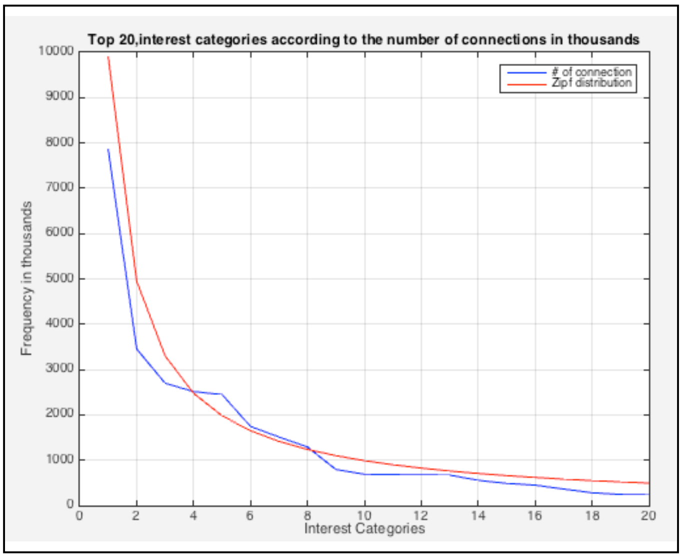


Fig. 2. Top 20, interest categories according to the number of connections (in thousands) and related Zipf's distribution

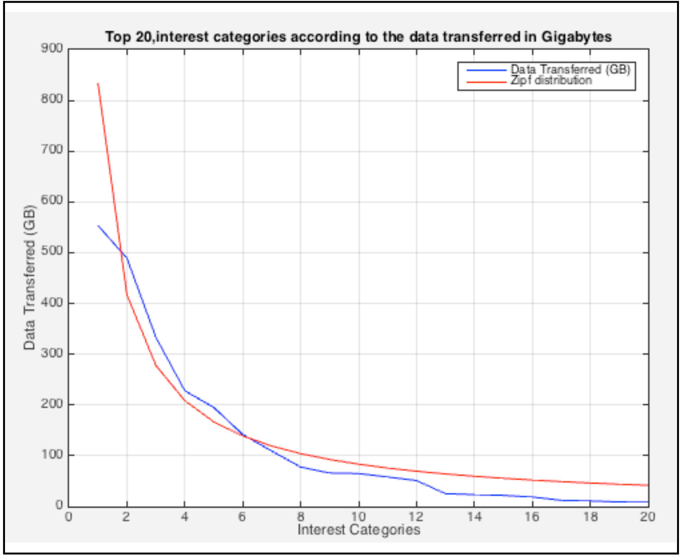


Fig. 3. Top 20, interest categories according to the data transferred (in Top 10 interest distributions according to the first preferences of subscribers)

The nature of Internet traffic pattern was analyzed by Zipf's law [13]. The main idea covered by the law can be adapted into our study as follows: the most popular sites generating incoming or outgoing traffic have the largest probability to be accessed. The total probability for N site is calculated as ln (N). The probability distribution for access ratio is as follows: The probability to access the most popular site is 1/ln(N), accessing to the second popular one is 1/2* ln (N), and so on. In addition to presenting the graph representation of table 1 and table 2, figure 2 and figure 3 also compares the Zipf's distribution with larger data set.

Tables 3- 5 present the results for Internet categories in which subscribers are assigned according to their Internet usage. The subscribers are assigned into interest categories according to their Internet usage characteristics. For a subscriber, the number of access within a category is divided by his/her total number of accesses and the results for each category group is calculated. After the calculations, the interest ratio vector is sorted in descending order to find the most popular interest category for that user.

TABLE III. TOP 10, INTEREST DISTRIBUTIONS ACCORDING TO THE FIRST PREFERENCES OF SUBSCRIBERS

Interest Categories	**Interest Ratio Frequencies and Ratios**		
	Ratio > 0.5	***Ratio > 0.35***	**Interest Categories**
Image Sharing	438	756	Image Sharing
Technology and Computer	98	229	Technology and Computer
Search Engines	44	109	Search Engines
Unknown	25	53	News
Games	23	51	Unknown
News	18	38	Games
Content Delivery Networks (CDN)	11	31	Shopping
Shopping	9	18	Content Delivery Networks (CDN)
Social Networks	7	16	Online Video/Audio
Online Video/Audio	7	13	Mobile

TABLE IV. TOP 10, INTEREST DISTRIBUTIONS ACCORDING TO THE SECOND PREFERENCES OF SUBSCRIBERS

Interest Categories	**Interest Ratio Frequencies and Ratios**		
	Ratio > 0.3	***Ratio > 0.2***	**Interest Categories**
Technology and Computer	31	163	Technology and Computer
Search Engines	17	105	Search Engines
Image Sharing	17	98	Image Sharing
Unknown	14	69	Unknown
Content Delivery Networks (CDN)	5	47	News
Social Networks	4	41	Advertisements
Advertisements	3	28	Shopping
News	3	26	Content Delivery Networks (CDN)
Shopping	3	13	Social Networks
Government and Organizations	2	12	Mobile

Table 3 presents the number of subscriber assignments according to their first preferences. The experiments for this category employed twice where the threshold values are set 0.5 and 0.35 respectively. Almost half of the subscribers had a choice with an interest ratio higher than 0.35. Interestingly, when the frequency ratio changed the popularity order of the interests changed. Same experiments repeated for second and third most popular interest categories and the results are summarized in Tables 4 and 5, respectively. The experiments are repeated with changing frequency ratios for both experiments.

TABLE V. TOP 10, INTEREST DISTRIBUTIONS ACCORDING TO THE THIRD PREFERENCES OF SUBSCRIBERS

Interest Categories	Interest Ratio Frequencies and Ratios		
	Ratio > 0.2	*Ratio > 0.1*	**Interest Categories**
Content Delivery Networks (CDN)	8	282	Advertisements
Technology and Computer	5	225	Search Engines
Image Sharing	4	192	Technology and Computer
Unknown	4	150	Unknown
Search Engines	4	97	News
Advertisements	3	96	Image Sharing
Classifieds and Auctions	1	73	Content Delivery Networks (CDN)
Online Storage	1	46	Shopping
Translation/Dictionary	1	38	Mobile
Government and Organizations	1	34	Online Video/Audio

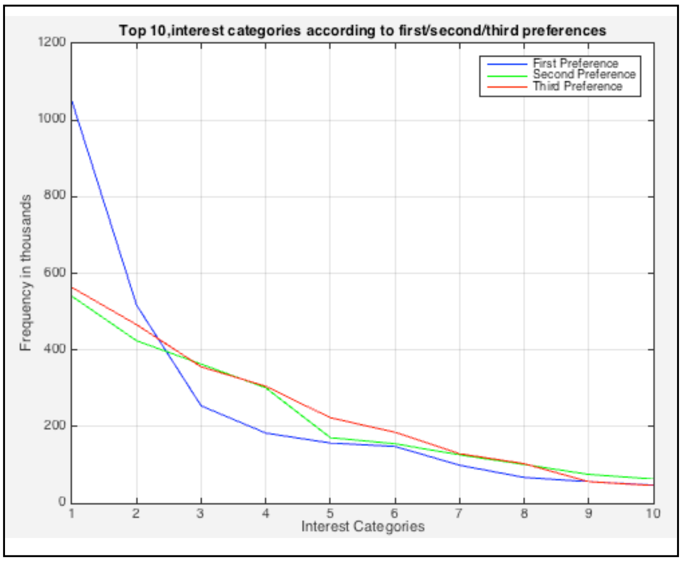


Fig. 4. Top 10, interest category subscriber frequencies according first/second/third preferences

Figure 4 presents Tables 3-5 together in graphical form. While second and third preferences have similar behaviors that starts from small values and decreases smoothly, 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 can only capture the difference. This reality has been observed during the calculations performed. Top 25 domains were excluded from the interest classifications in order to make difference between subscribers such as: google.com, mail.ru, apple.com, instagram.com. In addition to the famous sites, search engines and uncategorized domains are also excluded from the database.

After performing the purification steps the number of subscribers decreased to 2811. The number of total page visits during the experiments decreased to 828448, however, the number of connections for different objects decreased to 13426026. Although, the decrease in most of the parameters excluding the number of subscribers is about 40%, the total traffic consumed by the subscribers aggressively decreased to 628.39 GB. Naturally, the number of unique domains accessed decreased to 31811. The second interesting result is observed when the traffic of popular domains has been investigated. The traffic consumption ratio for top 10, 100, and 1000 domains decreased to 22.98%, 71.42%, and 95.03% respectively.

TABLE VI. TOP 10 INTEREST CATEGORIES ACCORDING TO THE NUMBER OF CONNECTIONS AFTER PURıFıCATıON

Interest Categories	Connection Information	
	# of Connections	*Connection ratio*
Technology and Computer	2091681	15.58%
News	1897084	14.13%
Shopping	874603	6.51%
Content Delivery Networks (CDN)	803592	5.99%
Business Services	653905	4.87%
Online Video/Audio	638528	4.76%
Games	551563	4.11%
Social Networks	485948	3.62%
Internet Portals	795062	3.43%
Marketing	690436	3.35%

TABLE VII. TOP 10 INTEREST CATEGORIES ACCORDING TO THE DATA TRANSFERRED (IN GIGABYTES) AFTER PURıFıCATıON

Interest Categories	Transferred Data Information	
	Data transferred (GB)	*Transfer Percentage*
Technology and Computer	98.73	15.71%
Online Video/Audio	92.89	14.78%
Content Delivery Networks (CDN)	72.21	11.49%
Pornography	58.15	9.25%
News	46.69	7.43%
Online Storage	46.47	73.95%
Games	35.60	5.67%
Internet Portals	23.25	3.70%
Shopping	12.70	2.02%
Education	12.22	1.95%

Tables 6 and 7 are purified counterpart of Tables 1 and 2. One of the most important finding in the experiments was having more divergent distribution on the interest categories according to the number of connections and the traffic incurred for the pages. Besides, differentiating characteristics are observed by having different interest categories like Education, Marketing, and Pornography. Although there is diversity in the interest category distribution, the nature adapts itself to Zipf's distribution after excluding a few initial samples as can be seen from Figures 5 and 6.

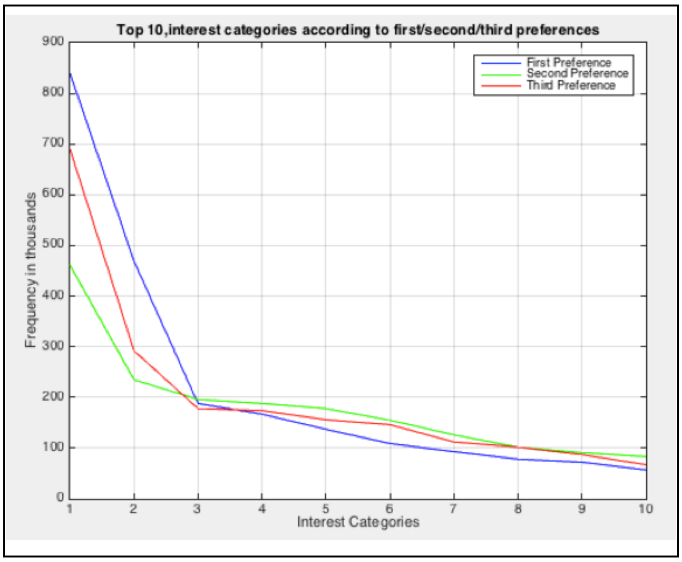


Fig. 5. Top 20, interest categories according to the number of connections (in thousands) and related Zipf's distribution after purification

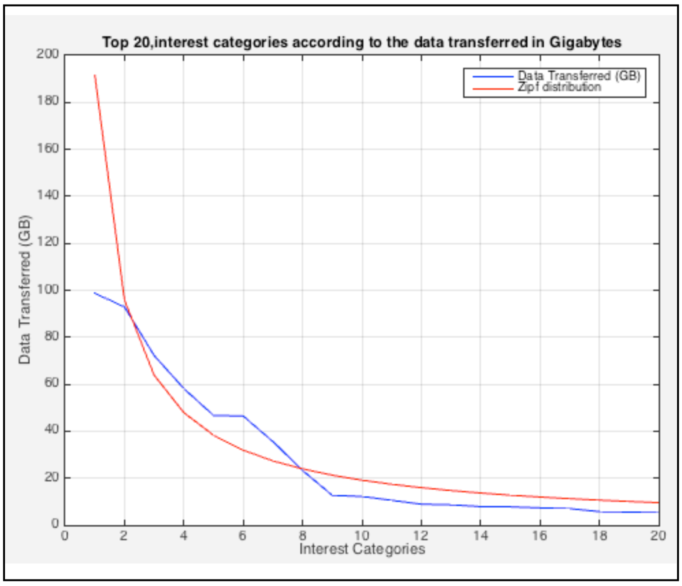


Fig. 6. Top 20, interest categories according to the data transfrred (in gigabytes) and related Zipf's distribution after purification

Tables 8-10 demonstrate the results of interest category assignment vectors after purification of the common traffic. On contrary to the tables 3-5, the results presents the diversity of the values obtained from the experiments which is the preferred behaviors to run with an advertisement network. Figure 7 also confirms the results of tables 8-10 when compared with figure 4 as well.

TABLE VIII. TOP 10, INTEREST DISTRIBUTIONS ACCORDING TO THE FIRST PREFERENCES OF SUBSCRIBERS AFTER PURIFICATION

Interest Categories	**Interest Ratio Frequencies and Ratios**		
	Ratio < 0.1	***Ratio > 0.35***	**Interest Categories**
Technology and Computer	336	86	Technology and Computer
News	152	35	News
Mobile	89	31	Games
Shopping	63	13	Content Delivery Networks (CDN)
Content Delivery Networks (CDN)	58	11	Shopping
Internet Portals	44	10	Online Video/Audio
Business Services	27	9	Mobile
Social Networks	23	8	Education
Online Video/Audio	22	7	Pornography
Games	21	7	Social Networks

TABLE IX. TOP 10, INTEREST DISTRIBUTIONS ACCORDING TO THE SECOND PREFERENCES OF SUBSCRIBERS AFTER PURIFICATION

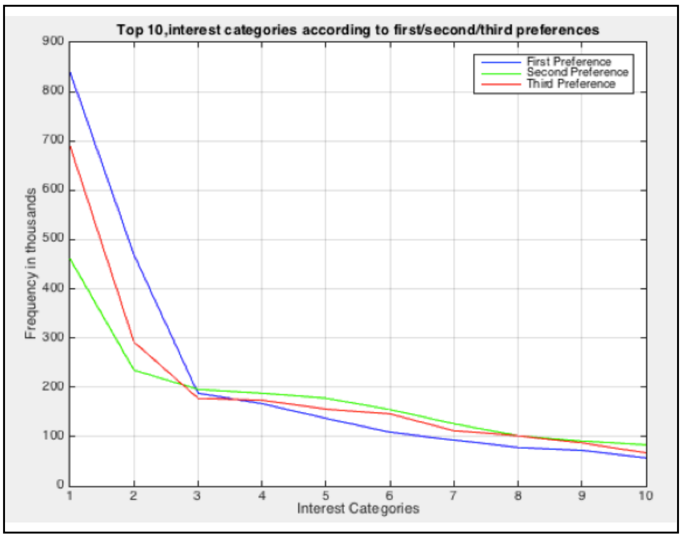
Interest Categories	**Interest Ratio Frequencies and Ratios**		
	Ratio < 0.1	***Ratio > 0.2***	**Interest Categories**
Technology and Computer	548	15	Technology and Computer
News	216	10	News
Mobile	144	5	Content Delivery Networks (CDN)
Content Delivery Networks (CDN)	136	4	Government and Organizations
Shopping	123	3	Mobile
Business Services	122	3	Shopping
Internet Portals	95	2	Sports
Online Video/Audio	79	2	Games
Sports	73	2	Business Services
Images CDN-Dynamic Content	57	2	Images CDN-Dynamic Content

Although the calculation of the interest vectors was the main argument of the system, in reality, it does not mean anything to the CSPs. The aim of the operators is to increase the revenues generated through value added services from customers. In order to simulate the effectiveness of the system, a test scenario has been generated. It was assumed that clicking an advertisement is 10% more likely if the content of the advertisement was in the same category of a subscriber having a first interest category score of 0.35. Similar assumptions can be done for buying decisions to market a new mobile phone. The manufacturer would query the subscribers having common interest in "Technology and Computer", "Mobile" and "Shopping" to show the advertisement of the new mobile phone rather than a subscriber having no interest on the above topics.

TABLE X. TOP 10, INTEREST DISTRIBUTIONS ACCORDING TO THE THIRD PREFERENCES OF SUBSCRIBERS AFTER PURIFICATION

Interest Categories	**Interest Ratio Frequencies and Ratios**		
	Ratio < 0.1	***Ratio > 0.2***	**Interest Categories**
Technology and Computer	548	15	Technology and Computer
News	216	10	News
Mobile	144	5	Content Delivery Networks (CDN)
Content Delivery Networks (CDN)	136	4	Government and Organizations
Shopping	123	3	Mobile
Business Services	122	3	Shopping
Internet Portals	95	2	Sports
Online Video/Audio	79	2	Games
Sports	73	2	Business Services
Images CDN-Dynamic Content	57	2	Images CDN-Dynamic Content

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



V. CONCLUSIONS AND FUTURE WORK

In this study, the effect of categorization of the subscribers for ad networks was investigated. Unlike conventional ad networking systems, the subscribers were categorized according to developed system compatible with IABs Tier 1 and Tier 2 categories. The aim of the study was to find differentiating points on subscribers to advertise them appropriate products to make additional revenue from the advertisement systems. 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 offered is rooted from the nature of Internet: The encrypted traffic. There is no unique and direct solution to the encrypted traffic to get information from.

In future, besides from categorization using web access logs, the system can be improved by using the information extracted from the applications such as gaming and social media. Additionally, the TSL protocol is a border on the usage of information extracted from search keys and social media. The possibility of the information retrieval from such systems will be investigated. Since the experiments conducted throughout the study is performed on the trace data of Internet access mainly during the daytime, the behaviors of the subscribers can change with time. The traffic pattern for the same subscriber would differ in different timeframes of the day and even different days in a week. One more issue to be covered would be investigation to find the correlation/diversity of mobile subscribers and landline subscribers.

REFERENCES

[1] Portio Research Ltd. Portio Research Mobile Factbook 2013, Report accessed online on August 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) January 26-27, 2013 Hong Kong (China)

[3] K., Grishikashvili, S. ,Dibb, M., Meadows, “Investigation into Big Data Impact on Digital Marketing”, Online Journal of Communication and Media Technologies Special Issue, October 2014.

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

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

[6] The Canadian Internet,, 2015. Resources accessed online on August 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”, Proceedings of the 1982 Academy of Marketing Science (AMS) Annual Conference Part of the series Developments in Marketing Science: Proceedings of the Academy of Marketing Science pp 519-524.

[8] D., Johansen, K., Friis, E.,Skovenborg and M., Grønbæk, “Food buying habits of people who buy wine or beer: cross sectional study”, BMJ. 2006 Mar 4; 332(7540): 519–522..

[9] J.,Lescovec, A.,Rajaraman, J.,Ullman,” Mining of Massive Datasets”, retrieved from http://www.mmds.org on August 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”, arXiv:1502.02215v1, February 2015.

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

[14] E.,Turban, D.,King, J.,K.,Lee, T.,P.,Liang, and D.,C.,Turban, “Social Commerce: Foundations, Social Marketing, and Advertising“, Electronic Commerce: A Managerial and Social Networks Perspective”, Part of the series 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, 2014, pp 282-305.