

ESTIMATION OF WEB TRAFFIC GENERATED BY USERS IN HOME NETWORKS

Roman A. Dunaytsev, Andrey V. Krendzel, Yevgeni A. Koucheryavy, Jarmo J. Harju
Institute of Communication Engineering, Tampere University of Technology
P.O. Box 553, SF-33101, Tampere, Finland
dunaitsev99@mail.ru, {krendzel, yk, harju}@cs.tut.fi

Abstract

Providing an access to the Internet and support of high-speed data transfer nowadays are the key issues for the most telecommunication operators. One of the perspective trends today is providing an Internet access for home users over home (campus) networks. At the stage of planning and preparation of system project for deployment of a network it is necessary to estimate traffic volume which will be generated by Web users in this network. Unfortunately a large diversity of Web content and Web users' behavior make the estimation problem quite complex. In this paper we present estimation method of Web traffic volume based on a probabilistic model of events initiated by Web users in a home network. It is proposed to make both a decomposition of Web content into some types and a distribution of Web users into some groups in accordance with a demand for different Web content. This approach allows segregating the common Web traffic into several segments and calculating the data traffic generated by different types of Web users in a home network while requesting different types of Web content.

Key Words

Web traffic estimation, user behavior.

1. Introduction

According to statistical data home users form a significant part from the total number of Internet users. There is a variety of Internet access methods for home users: dial-up, ISDN, xDSL and others. One of the perspective trends today is an Internet access over home (campus) networks. In this case computers of home users are interconnected into a local area network (LAN) so that they are able to communicate, exchange information and share resources. At the stage of preparation of system project for deployment of a network one of the arising problems is how to connect it to an Internet gateway, i.e. to connect it by ADSL, dedicated line and so on. Among different

parameters which have to be taken into account the volume of Web traffic generated by users of a home network is quite important factor.

There is a large diversity of Web content and Web users' behavior. The examples of Web users' activities are Web-surfing, chat, downloading pictorial, audio and video files and so on. As a result, estimation problem of the data traffic generated by Web users becomes quite complex when planning home networks. In this paper the estimation method of the data traffic generated by Web users in home networks is considered. It is based on a probabilistic model of events initiated by Web users. It is suggested to perform both a decomposition of Web content into some types and a decomposition of potential Web users into some groups. This method is based on the approach proposed in [1].

The structure of the paper is organised as follows. We start with the decomposition of Web content into some types. An amount of requests and an amount of transferred data per request are used as criterion for such decomposition. Then the distribution of Web users into groups is carried out taking into account a non-uniform demand for different Web content in accordance with the Pareto law. So the probabilistic model of events initiated by Web users in home network comprises two exhaustive classes of statistically independent events. It allows segregating the common Web traffic into several segments. After this we calculate the rate of requests and determine the data traffic generated by Web users in a home network. Finally, we present a case study using obtained expressions.

2. Decomposition of Web content into types

One of the main features of data traffic generated by Web users is a significant diversity because of different types of informational content provided by World Wide Web (WWW) to users. Hence, it is worthwhile to make a decomposition of Web content into several types in such a way that characteristics of the generated traffic, while requesting one of content types, would be approximately

identical. In this paper we present an approach for the decomposition of Web content into three types $\{i; i=1,2,3\}$. Note that there are no special limitations for the choice of a number of types.

The first type of Web content $i=1$ includes text pages with a small amount of graphics, E-mail (with an access from a browser), chat and so on. This type is denoted as *pages*. A demand for this type of content among users is high, however an average volume of transferred data while requesting this type of content by a user is quite low, from tens to hundreds of Kbytes.

The second type $i=2$ provides an opportunity for users to download files such as jpeg, gif, doc, pdf, zip and etc. This type is denoted as *pictures*. An average volume of transferred data while requesting this type of content by a user may be characterized as the mean level, hundreds of Kbytes.

The third type $i=3$ deals with multimedia data, i.e. audio and video files (mp3, mpeg, avi and so on). This type is denoted as *multimedia*. The volume of transferred data while requesting this type of content by a user may achieve several tens and hundreds of Mbytes.

The initial information about specific (per each type) distribution of the total amount of requests in the hour with the maximum number of active Web users $\{\gamma_i; i=1,2,3\}$ may be used as the numerical criterion for the decomposition of Web content into three types. These initial parameters form the exhaustive set and may be denoted as

$$\gamma_i, \quad i=1,2,3, \quad \gamma_1 + \gamma_2 + \gamma_3 = 1, \quad (1)$$

where $i=1,2,3$ corresponds to *pages*, *pictures*, *multimedia* types of Web content.

Note that here and after under request we understand opening by a user a Web page, clicking on a hyperlink, downloading some file and so on.

Thus the result of the submitted decomposition is the splitting of Web content into three types. Such approach has the following advantage. Web users while requesting each type will generate traffic with variance of values of parameters substantially smaller than one relating to the total Web traffic.

3. Decomposition of Web users into groups

A demand for different Web content mainly depends on both user's interests and tariffs on Internet access. Let's consider a case when charging in a home network is fulfilled according to volume of transferred data. Then fees for downloading information from different types of Web content will be unequal. As a result, a non-uniformity of the demand for different types of Web content will take place among Web users. For this case it is worthwhile to perform a decomposition of Web users into groups by their level of income.

Note that in a case of unlimited access (fees don't increase with volume of transferred data) similar approach can be used for the decomposition of Web users into groups by their interests and needs for different Web content.

Usually a non-uniform distribution of incomes between inhabitants is characterized by the Gini coefficients [2], [3]. Values of the Gini coefficients may be defined on a basis of statistical information and marketing research regarding the demand for different Web content.

Let the non-uniformity of the demand for each type of Web content be defined as

$$G_i, \quad i=1,2,3, \quad (2)$$

where G_i is the Gini coefficient for the i -th type of Web content.

Let's assume that the non-uniformity of the distribution of the demand for different types of Web content corresponds to the Pareto law [4], [5] and [6]. Then parameters of the Pareto distribution may be calculated as follows [2], [6]

$$\alpha_i = \frac{G_i + 1}{2G_i}, \quad i=1,2,3. \quad (3)$$

It is obviously that the least non-uniformity of the distribution will take place for the *pages* type and the largest one will take place for the *multimedia* type, i.e.

$$\alpha_1 > \alpha_2 > \alpha_3 > 1. \quad (4)$$

Then it is possible to distribute Web users into groups in accordance with their demand for different Web content when given (3) and (4). As a rule for the decomposition of Web users into three groups $\{j; j=1,2,3\}$ the following one described below may be applied. Note that there are no special limitations for the choice of a number of groups.

The wealthiest and the most interested in multimedia information users producing 80% of requests for the *multimedia* type of Web content are included into the third group $j=3$, denoted as *heavy*. The Lorenz curves corresponding to the Pareto distribution with parameter α may be written as follows [7], [8]

$$Q(\alpha, x) = 1 - (1 - F(x))^{\frac{\alpha-1}{\alpha}}. \quad (5)$$

Using the Lorenz curves (Fig. 1) it is possible to determine the relative number of users in the *heavy* group F_3 as

$$F_3 = 0.8^{\frac{\alpha_3}{(\alpha_3-1)}}. \quad (6)$$

Let users from the third group and the second group create 80% of requests for the *pictures* type of Web content. Then the relative number of users in the second group $j=2$, denoted as *medium*, may be determined as

$$F_2 = 0.8^{\frac{\alpha_2}{(\alpha_2-1)}} - F_3. \quad (7)$$

The relative number of users in the first group $j=1$, denoted as *light*, may be found as

$$F_1 = 1 - F_2 - F_3. \quad (8)$$

The expressions (6), (7) and (8) give the rule of a definition of three groups of Web users in accordance with their demand for different types of Web content defined in the section 2.

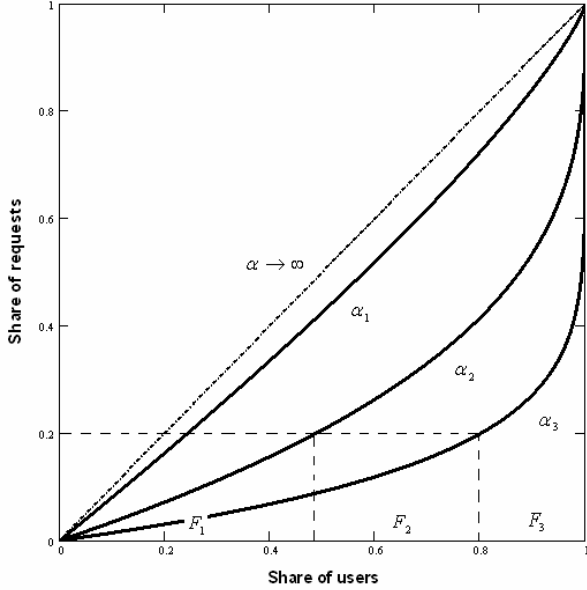


Fig. 1. Decomposition of Web users into groups.

4. Calculation of the rate of requests

The decomposition of Web content and the decomposition of Web users give a possibility to form a probabilistic model of the initiation of requests based on intersection of events from two statistically independent exhaustive classes.

The events included in the first class are denoted by the index $i = 1, 2, 3$. They correspond with user's request for one of the types of Web content (*pages*, *pictures* and *multimedia* types respectively). The events included in the second class are denoted by the index $j = 1, 2, 3$. They correspond with request, initiated by user from one of the groups (*light*, *medium* and *heavy* groups respectively). The first class and the second class of events are exhaustive by a definition since the expressions (1) and (8) are true. It is supposed that these classes of events are independent, in other words, the probability of intersection of the events are equal to the product of probabilities each of the events.

Thus, the above-mentioned decomposition of Web content into three types and the above-mentioned decomposition of Web users into three groups allow us to define nine events. The events form the exhaustive set and are denoted below by the dual index i, j ; $i = 1, 2, 3$, $j = 1, 2, 3$. The first event ($i = 1, j = 1$) is that a user from the first group ($j = 1$)

initiates a request for data from the first type of Web content ($i = 1$). At the second event ($i = 2, j = 1$) a user from the first group initiates a request for data from the second type of Web content and so on.

Thus the probabilistic model of the events allows us to segregate nine segments from the total Web traffic. A variance of values of parameters in each segment will be less than one relating to the total Web traffic.

So the problem is to determine average values of the rate of requests. We consider the approach for the calculation of the specific (per a user) rate of requests λ_{ij} $\{i, j = 1, 2, 3\}$ in the hour with the maximum number of active Web users for nine intersections of events from two above-mentioned classes. It is based on solving for a system of three equations that should be formed for each type of Web content.

Each system of the equations may be assigned on a basis of values β_{ij} , where β_{ij} is a share of requests in the hour with the maximum number of active Web users relating to users of the j -th group when the i -th type of content is requested.

Values β_{ij} should be calculated for each of nine events at given α_i , $i = 1, 2, 3$, F_j , $j = 1, 2, 3$. Taking into account the expressions (3), (5), (6), (7) and (8) and that $\sum_j \beta_{ij} = 1$, $i, j = 1, 2, 3$ the following systems of the equations may be defined

$$\begin{cases} \beta_{i1} = 1 - (1 - F_1)^{\frac{(\alpha_i - 1)}{\alpha_i}} \\ \beta_{i2} = 1 - (1 - F_1 - F_2)^{\frac{(\alpha_i - 1)}{\alpha_i}} - \beta_{i1} \\ \beta_{i3} = 1 - \beta_{i1} - \beta_{i2} \end{cases} \quad i = 1, 2, 3 \quad (9)$$

These systems of the equations allow calculating the distribution of requests for three groups of Web users taking into account the non-uniformity of a demand of different types of Web content.

Using the parameter λ_{ij} $\{i, j = 1, 2, 3\}$ defined earlier it is quite easy to determine the expression for the total rate of requests in the hour with the maximum number of active Web users that generated by all users of the j -th group when they requested content from the i -th type of content as $NkF_j\lambda_{ij}$ $\{i, j = 1, 2, 3\}$, where N is the total number of users in a home network, k is a share of active users in the hour with the maximum number of active Web users.

Then a share of requests in the hour with the maximum number of active Web users relating to users of the j -th group when the i -th type of content is requested may be determined as

$$\beta_{ij} = \frac{F_j \lambda_{ij}}{\sum_j \lambda_{ij} F_j}, \quad i, j = 1, 2, 3. \quad (10)$$

The expressions (9) and (10) give three systems of linear equations $\{i=1,2,3\}$ with three unknown λ_{ij} $\{i, j = 1, 2, 3\}$ in each system.

$$\begin{cases} (\beta_{11}-1)F_1\lambda_{11} + \beta_{11}F_2\lambda_{12} + \beta_{11}F_3\lambda_{13} = 0 \\ \beta_{21}F_1\lambda_{11} + (\beta_{22}-1)F_2\lambda_{22} + \beta_{22}F_3\lambda_{23} = 0 \\ \beta_{31}F_1\lambda_{11} + \beta_{32}F_2\lambda_{22} + (\beta_{33}-1)F_3\lambda_{33} = 0 \end{cases} \quad i=1,2,3. \quad (11)$$

These systems may be presented in vector-matrix form as follows

$$\begin{pmatrix} (\beta_{11}-1)F_1 & \beta_{11}F_2 & \beta_{11}F_3 \\ \beta_{21}F_1 & (\beta_{22}-1)F_2 & \beta_{22}F_3 \\ \beta_{31}F_1 & \beta_{32}F_2 & (\beta_{33}-1)F_3 \end{pmatrix} \otimes \begin{pmatrix} \lambda_{11} \\ \lambda_{22} \\ \lambda_{33} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad i=1,2,3 \quad (12)$$

or, more compactly

$$\mathbf{A}_i \boldsymbol{\lambda}_i = \mathbf{0}, \quad i = 1, 2, 3. \quad (13)$$

The rank of matrix \mathbf{A}_i is equal to 2 and equal to the rank of augmented matrix. Therefore the linear equations in these systems are simultaneous. These systems of equations are indefinite (since the number of linear equations more than the rank of matrix \mathbf{A}_i) and homogeneous (since all absolute terms are equal to 0). So these systems of equations have infinite number of solution sets.

Let's determine solution set when $i = 1$. For this we can select, for example, the first and the third linear equations. Then

$$\begin{aligned} \lambda_{11} &= \begin{vmatrix} \beta_{11}F_2 & \beta_{11}F_3 \\ \beta_{13}F_2 & (\beta_{13}-1)F_3 \end{vmatrix} \otimes t = -\beta_{11}F_2F_3t, \\ \lambda_{12} &= \begin{vmatrix} \beta_{11}F_3 & (\beta_{11}-1)F_1 \\ (\beta_{13}-1)F_3 & \beta_{13}F_1 \end{vmatrix} \otimes t = -\beta_{12}F_1F_3t, \\ \lambda_{13} &= \begin{vmatrix} (\beta_{11}-1)F_1 & \beta_{11}F_2 \\ \beta_{13}F_1 & \beta_{13}F_2 \end{vmatrix} \otimes t = -\beta_{13}F_1F_2t, \end{aligned} \quad (14)$$

where t is an arbitrary real number.

In order to concretize the solutions (14) it is necessary to add to the input data the value λ_{11} that may be estimated on a basis of statistical information. Then using (14) we obtain

$$\begin{aligned} t &= -\frac{\lambda_{11}}{\beta_{11}F_2F_3}, \\ \lambda_{12} &= \frac{\lambda_{11}\beta_{12}F_1}{\beta_{11}F_2}, \\ \lambda_{13} &= \frac{\lambda_{11}\beta_{13}F_1}{\beta_{11}F_3}. \end{aligned} \quad (15)$$

In order to determine solution sets when $i = 2, 3$ it is necessary to use input data defined in (1), i.e. values of the parameter γ_i , $i = 1, 2, 3$, then

$C\gamma_j = Nk \sum_j \lambda_{ij} F_j$, where $i, j = 1, 2, 3$, C is the total number of requests in the hour with the maximum number of active Web users. Hence

$$\gamma_2 \sum_j \lambda_{1j} F_j = \gamma_1 \sum_j \lambda_{2j} F_j, \quad (16)$$

$$\gamma_3 \sum_j \lambda_{1j} F_j = \gamma_1 \sum_j \lambda_{3j} F_j.$$

From here

$$\sum_j \lambda_{2j} F_j = \frac{\gamma_2 \sum_j \lambda_{1j} F_j}{\gamma_1}, \quad (17)$$

$$\sum_j \lambda_{3j} F_j = \frac{\gamma_3 \sum_j \lambda_{1j} F_j}{\gamma_1}.$$

Taking into account the obtained sum $\sum_j \lambda_{1j} F_j$ and expressions (10) and (17), we have the rest two solution sets when $i = 2, 3$ as

$$\lambda_{ij} = \frac{\gamma_i \beta_{ij} \sum_j \lambda_{1j} F_j}{\gamma_1 F_j}, \quad i = 2, 3, j = 1, 2, 3. \quad (18)$$

Thus we have obtained the estimations (14) and (18) of the rate of requests in the hour with the maximum number of active Web users for eight segments of the data traffic generated by Web users. The value of the first segment ($i, j = 1$) is included in the input data.

5. Calculation of traffic volume generated by Web users

In order to calculate the traffic volume generated by Web users it is necessary to add to the input data the values of an average amount of transferred data per request to the i -th type of Web content. These parameters are denoted as

$$w_i, \text{KB/request}, \quad i = 1, 2, 3. \quad (19)$$

The values of w_i were received by empirical way. To analyze traffic generated by requesting for different types of Web content we investigated traffic traces gathered with the help of network monitor and protocol analyzer – ColaSoft Capsa 3.0 [9]. We used a packet trace-based approach because it allowed us to capture the behavior of individual users. Obtained results are presented in Table 1 and Fig. 2–4.

The total traffic volume generated by Web users during the hour with the maximum number of active Web users while requesting for i -th type of Web content may be defined as

$$S_i = Nk w_i \sum_j \lambda_{ij} F_j, \quad \text{KB}, \quad (20)$$

and, at last, the final expression for the total traffic volume generated by Web users during the hour with the maximum number of active Web users is

$$S = Nk \sum_i w_i \sum_j \lambda_{ij} F_j, \quad KB. \quad (21)$$

The presented expressions have been obtained when both the number of types of Web content and the number of groups of Web users are equal to three. However, as it has been emphasized above this method may be used in cases when the values of $\{i, j = 1, 2, 3\}$ are more than three.

Table 1

Parameter	Type of Web content, i		
	<i>pages,</i> $i = 1$	<i>pictures,</i> $i = 2$	<i>multimedia,</i> $i = 3$
Traffic volume, MB	284,40	307,55	453,44
Number of requests to the i -th type of Web content	2042	1568	116
Average amount of data per request to the i -th type of Web content (W_i), KB/request	140	200	4000

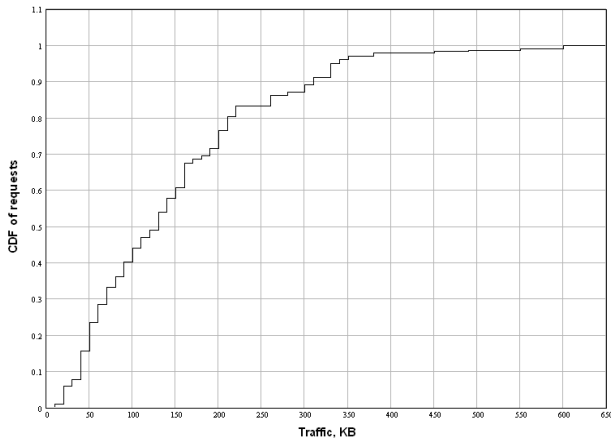


Fig. 2. The empirical cumulative distribution function (CDF) of requests to the *pages* type of Web content as a function of transfer size.

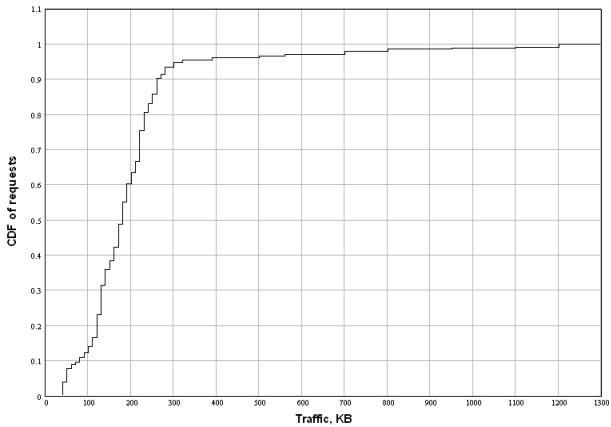


Fig. 3. The empirical cumulative distribution function (CDF) of requests to the *pictures* type of Web content as a function of transfer size.

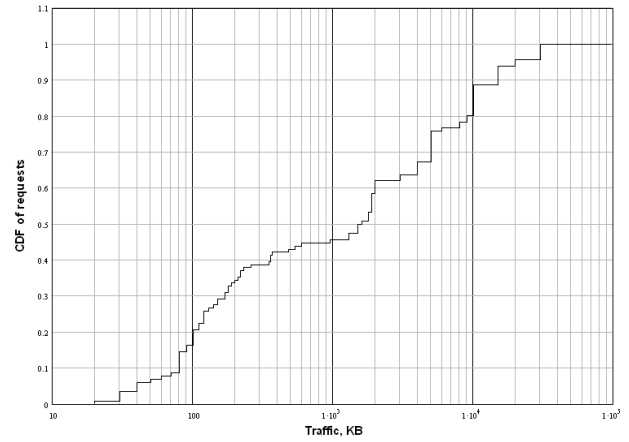


Fig. 4. The empirical cumulative distribution function (CDF) of requests to the *multimedia* type of Web content as a function of transfer size.

6. Case study

Let's consider an example of calculation of traffic volume generated by Web users. The initial values are presented in Table 2.

The calculations have been made using the approach and the expressions presented in previous sections. The main results are the follows:

- total Web traffic generated by 50 users of home network during the hour with the maximum number of active Web users is about 1 GB;
- traffic, generated by users while requesting to *pages* type of content is 169 MB, while requesting to *pictures* type is 125 MB, while requesting to *multimedia* type is 724 MB.

Table 2

Parameter	Type of Web content, i			Meaning
	<i>pages,</i> $i = 1$	<i>pictures,</i> $i = 2$	<i>multimedia,</i> $i = 3$	
α_i	3,5	1,5	1,16	Pareto distribution coefficient
γ_i	0,6	0,31	0,09	Share of requests to the i -th type of Web content in the hour with the maximum number of active Web users
W_i , KB/request	140	200	4000	Average amount of data per request to the i -th type of Web content
N	50			Number of users
k	0,6			Share of active users in the hour with the maximum number of active Web users
λ_{11} , request/user	35			Rate of requests to the first type of Web content from the first group of Web users in the hour with the maximum number of active Web users

7. Conclusion

In this paper the estimation method of Web traffic generated by users in a home networks has been presented. The use of this method allows making easier both the preparation of business-plans and system projects for deployment of home networks.

The definition of the proper values of the input data requires the further study and can be received on basis of statistical research.

References:

- [1] A. Krendzel, Y. Koucheryavy, S. Lopatin, J. Harju, Estimation method for data traffic generated by 3G users, *Proc. ICC-2004 IEEE International Conf. on Communications*, Paris, France, 2004.
- [2] M. Kendall, A. Stuart, *The advanced theory of statistics. Distribution theory* (London: C. Griffin & Co., 1962).
- [3] C.V. Brown, P.M. Jackson, *Public sector economics* (Oxford: Blackwell, 1990).
- [4] P. Hardwick, B. Khan, J. Langmead, *An introduction to modern economics* (London: Longman, 1994).
- [5] E. Gumbel, *Statistics of extremes* (New York: Columbia University Press, 1962).
- [6] L.E. Varakin, The Pareto law and the rule 20/80: the distribution of incomes and telecommunication services, *MAC proceedings*, 1, 1997, 3–10.
- [7] C. Dagum, The generation and distribution of income, the Lorenz curve and the Gini ratio, *Economie Appliquée*, 33, 1980, 327–367.
- [8] P.M. DeRusso, R. Roy, C. Close, A. Desrochers, *State variables for engineers* (NY: John Wiley & Sons, 1998).
- [9] Colasoft Co., Ltd. <http://www.colasoft.com/>.