Method of calculating
the true number of real users of web sites

1. Introduction

One of the basic indicators of web sites performance measurement is the number of visitors (Internet users) who visited the web site under study in a given period of time. In order to properly estimate the value for this indicator, one should take into account a lot of different complex factors (unlike in the case of page views registered on the web site under study), stemming from Internet communication technology as well as from different ways of using the Internet by its users.

The hereby document describes the unique and reserved method of calculating the true number of visitors (real users) to the web sites being studied in a given period of time devised by the Gemius SA research company. This method makes use of findings obtained by gemiusTraffic site-centric system along with information on the size of the whole Internet users population.

2. Number of cookies vs. number of Internet users

All site-centric and adserver systems present the number of cookies registered on the web site under study as the indicator pointing to the number of visitors to the web site being monitored by the system. Unfortunately, this indicator, due to independent causes, does not take into consideration the phenomenon of cookie deletion which manifests itself in some users occasionally removing (either on purpose or accidentally) the cookies stored in their computers. In such a way, those users are detected multiple times within the researching period by the system. The phenomenon of cookie deletion has a crucial impact on the fact that the number of registered cookies (it proves to be true especially for longer periods of time) significantly exceeds the true value for the number of visitors. If we take, for instance, a 30-day period, any site-centric and adserver system will recognize an Internet user who keeps deleting cookies from his/her computer on everyday basis as thirty different users!

When estimating the true number of the users who visited the web site under study, one should additionally take into account other factors which influence the fact that the number of registered cookies does not correspond to the number of the real users. The extent of significance of possible impact of those instances when one computer with one user's profile (one cookie) is used by several people as well as when a few computers are used by a single person is currently under investigation. However, at present it may already be stated that information on the size and socio-demographic structure of Internet users population allows for proper estimation of the real number of unique visitors to the web site being studied.

3. Gemius algorithm for calculating real users

In order to estimate the number of real users of the web site within a given period of time (let it be denoted by U_W) what one should do at first is to estimate the number of cookies which the system registered on the web site being monitored as if there were no such thing as cookie

deletion at all. Then, one should estimate the reach of the web site and finally come to the output value for the users of the web site.

The calculation procedure looks as following:

a) One should calculate the number of page views done by all the visitors to the web site – let it be denoted by O_W.

b) Then, one should calculate the number of those cookies about which they have reliable knowledge that the cookies have existed for the whole researching period (a cookie which has doubtless existed for the whole researching period is a cookie which existed before the researching period and persists beyond it) – let it be denoted by C_D.

c) Then, one should calculate the number of page views generated by the cookies specified in paragraph b) – let it be denoted by O_D.

d) Then, one should calculate the number of cookies which would be registerd on the web site under study if non of the the cookies were deleted, following the formula: $C_W = (O_W/O_D) * C_D$.

e) Similarly, taking into account the number of all web sites being studied instead of just the one in question, one should calculate the number of cookies which would be registered on those web sites if non of the cookies were deleted – let it be denoted by C_P.

f) Then, one should calculate the relative reach of the web site being studied, following the formula: $Z_W = C_W / C_P$.

g) If the size of the Internet users population among all web sites being studied within the researching period is denoted by P, then the formula for calculating users of the web site under study will be the following: $U_W = Z_W * P$.

When doing the whole calculation procedure, only the page views (consequently, also the cookies) generated on the computers with IP adresses coming from the domestic country should be taken into account.

The following assumptions serve as the basis for the above-outlined algorithm:

1. There is a possiblility to isolate the group of the cookies specified in the above paragraph b). One may monitor cookies remaining intact not only on the web site under study, but also on a possibly largest amount of web sites (the whole Internet network being the best source of information) as well as existing not only within the researching period but also before and after it (for the researching period lasting one month it would be suitable to track cookies remaining active in at least one preceding and one following months). What arises from this assumption is the fact that research results must be obviously published with some delay – the time is needed to verify whether a cookie belongs to the group of cookies described in paragraph b). In order to fulfill this assumption, it is crucial to be technically capable of monitoring and handling data on cookies remaining intact. gemiusTraffic research meets this requirement.
2. The cookies described in paragraph b) constitute a representative group for all cookies, especially, as far as the average of generated page views is concerned. In the case of all the research carried out by now, this assumption has been proved to be correct through comparative analyses of all possible bahavioural patterns of the representative group of cookies compared and contrasted with the rest of the cookies (there have been conducted analyses of i.a.: page views, visits, frequency of page

views and visits within longer periods of time, geographic location, operational systems, browsers, etc.).

3. The ratio of page views generated by the cookies specified in paragraph b) on all the web sites being monitored to the number of all page views generated on all the web sites being monitored equals the ratio of page views generated on all web sites (including those not being under study) by the same group of cookies to the number of all page views done on those web sites. This hypothesis has been proved to be correct through analyses of outcomes of panel studies in which the ratios of page views generated on web sites monitored to those not being monitored by the site-centric system performed by people deleting as well as by those not deleting cookie files who at least once visited minimum one of the web sites under study were taken into account.
4. The relative reach Z_W of the web sites under study estimated for a given period of time on the basis of the isolated group of cookies correctly estimates the reach of the web sites under study. This hypothesis is currently under in-depth examination, however, all the analyses conducted by now have lent credit to the assumption.
5. The size of Internet users population P properly estimates the number of all Internet users visiting all the web sites being studied over the researching period.

4. Summary

The above-described algorithm allows to estimate the real number of unique users of web sites under study. This algorithm takes into account the factors which may potentially lead to obtaining results deviating from reality such as: the possibility that one computer may be used by a few different people, the possibility that one person may use several computers/user's profiles and what is of utmost magnitude – the persistent phenomenon of cookie deletion.