# Tracking Activities of Real Persons in Weblogs

**Csaba Legány, Ferenc Kovacs**

Department of Automation and Applied Informatics
Budapest University of Technology and Economy
Goldmann Gy. tér 3, H-1111 Budapest, Hungary
{csaba.legany, ferenc.kovacs}@aut.bme.hu

*Abstract: It is very important for Internet content providers to keep track of the amount of visitors of their sites. Web auditing companies generate enormous weblogs from which allows the number of visitors to be estimated with the help of complex techniques. This paper introduces a system designed to process these weblog efficiently with special focus on identifying and separating the activities belonging to the same real persons. Compression and indexing methods are given how to preprocess the weblogs to derive further statistical data in a simpler and faster way.*

*Keywords: web auditing, system architecture, cookie chains, Internet user, real person*

## 1 Introduction

As the Internet became popular, the amount of users visiting web sites increased and the online media became an important part of the advertisement market. There are several web audit companies in Hungary as well [1, 2, 3] measuring the amount of visitors of web sites, this can be achieved by manual (questionnaires, survey, opinion poll) or by fully automated means (processing weblog [4, 5, 6, 7]).

Our paper focuses on this second area. Automated weblog processing has a much wider, and reliable sampling base, but is also harder to interpret. The advertisers are not primarily interested in the amount of downloads of a given site, in fact they would like to estimate how many different persons they can reach with their advertisements. It is impossible to measure the amount of real persons directly during the auditing process because such information is not available at the server when logging user downloads. In order to solve this problem, the number of Internet users is estimated, from which the number of real persons can be estimated with complex business logic. An Internet user is in fact no more than a triple of a computer, a user account and a web client (browser) currently in use while the downloading was done, which means that if a real person uses multiple computers / accounts / web clients, he or she will be logged as multiple Internet users [8, 9].

This paper introduces a system architecture that is able to support advanced methods for associating internet users with real persons in large weblogs. After the short introduction, Section 2 describes the structure of weblogs, while Section 3 detailes the weblog processing. The data files used by the system can be found in Section 4, while the system architecture with their layer functionalities are to be found in the Section 5. Section 6 summarizes our results.

## 2 Structure of Weblogs

The weblogs of auditing companies [1, 4] seek for including all the information needed to identify Internet users in their weblogs. The most common identification method is the use of C3 cookies (Third Party Cookies) [10], along with the use of C1 cookie (First Party Cookies) [11]. The main difference between the two types of cookies is that while C1 cookies are set by the page browsed currently, C3 cookies are placed by foreign, third party site. Since web pages often contain advertisements from foreign sites [5], they can also set C3 cookies on the clients but client-side security software might remove them periodically. The two cookie-based user identification methods are able to identify and trace Internet users; however they are unable to associate Internet users with real persons. A third identification technique is needed to solve this problem. Websites having registration database (e.g. e-mail providers) identify their users with special unique login identifiers (noted as MID-s). In order to protect the privacy of real persons, even MID-s do not reveal personal information but they can be used to separate real persons from each other [5, 6].

| Name | Size (byte) |
|---|---|
| C3 | 22 |
| C1 | 26 |
| MID | 40 |
| IP Address | 4 |
| Timestamp | 4 |
| UCode | 16 |
| Url | 100 |
| Web client | 180 |
| Summary | 392 |

Figure 1
Structure of Weblogs

Weblogs contain not only the three different user identifications but also the following data of Internet users: IP, Timestamp, Ucode (the topic code for the visited domain), URL, Web client. The web client contains not only the browser's name but also gives information about the client's operating system, browser plug-

ins, advertisement software installed, whether .NET Framework is installed, msn messenger version (if available) etc. Figure 1 summarizes the structure of weblogs subject to our investigations. The weblog is a logical object built up out of a sequence of log files. A file can contain a maximum of 5.000.000 pieces of 392 byte sized records, which means that their usual size is 1.9 GB (except the last fragment).

Other types of weblog might have slightly different format and fields to track the user activity. For example they can contain the following data:

- C3 cookie, date, browser type, navigation history and IP address [5]
- C3 cookie, date, user agent, url, uri, IP address, server protocol, request method [7]
- Cookies, IP address, username, date, file name and size, browser type, url, http status code [9].

# 3 Weblog Processing

Section 2 introduced three different user identification methods (C1 and C3 cookies and MID-s) and also described the detailed structure of weblogs. In order to process these weblogs to find the real persons behind Internet users efficiently, it is very important to separate and compress that data, which is required to join Internet users into real persons. These data are the following: C3 and C1 cookies and MID. Cookie Chains (noted as CC) are log entries and cookies belonging to a single Internet user. The CC-s can be joined by the MID-s to form cookie networks (noted as CN). The CC-s and CN-s have to be assigned to every log entry so that statistical calculations could be estimated on them. The following examples demonstrate the process of building CC-s and CN-s. Figure 2 depicts the C1, C3 and MID columns of a sample weblog.

| Entry | C1 | C3 | MID | Timestamp | IP | Client |
|---|---|---|---|---|---|---|
| 1 | A1 | B1 | | 01/09/2006 15h45’23.23 | 152.66.23.2 | IE |
| 2 | A1 | B1 | P1 | 01/09/2006 15h45’25.12 | 62.122.12.3 | IE |
| 3 | A43 | B32 | P1 | 01/09/2006 15h47’13.34 | 62.64.184.134 | Firefox |
| 4 | A43 | B33 | | 01/09/2006 15h47’13.35 | 62.64.184.135 | IE |
| 5 | A1 | B3 | | 01/09/2006 15h47’14.37 | 62.63.184.132 | IE |
| 6 | A34 | B14 | P1 | 01/09/2006 15h47’35.01 | 62.84.123.113 | Netscape |
| 7 | A35 | B14 | P2 | 01/09/2006 15h47’36.34 | 63.23.145.114 | IE |
| 8 | A56 | B87 | | 01/09/2006 15h47’43.12 | 63.45.122.151 | Netscape |
| 9 | A8 | B9 | | 01/09/2006 15h48’11.12 | 152.66.23.4 | IE |
| 10 | A1 | B9 | | 01/09/2006 15h49’10.32 | 81.23.124.122 | IE |

Figure 2
Sample weblog fragment

The following rules can be used to build a cookie chain from weblog entries:

- If two entries have the same C1 and C3 cookies, they belong to the same chain (the 1st and the 2nd entry in the example)
- If two entries have the same C1 cookies but are having different C3 cookies, they belong to the same chain. This happens if the C3 cookies were deleted on the client computer, the C3 cookies prevent chains to be split up in such cases (4th, 7th entry).
- If two entries have the same C3 cookies but different C1 cookies they belong to the same chain (6th and 7th entries).

The highlighted entries belong to one cookie chain (CC1):

- C1 cookies: A1, A8
- C3 cookies: B1, B3, B9
- MID-s: P1

Other cookie chains found in the sample weblog:

- CC2:
    - C1 cookies: A43
    - C3 cookies: B32, B33
    - MID-s: P1
- CC3:
    - C1 cookies: A34, A35
    - C3 cookies: B14
    - MID-s: P1, P2
- CC4:
    - C1 cookies: A56
    - C3 cookies: B87
    - MID-s: none

The following example demonstrates the merging of cookie chains to form cookie networks. Figure 3 displays a table containing four cookie chains and the cookie networks that can be built from them. The rule of building cookie networks is simple: if the chains have common MID-s they belong to the same network.

| CC | C1 | C3 | MID |
|---|---|---|---|
| CC1 | A1,A8 | C1,C3,C9 | P1 |
| CC2 | A43 | B32, B33 | P1 |
| CC3 | A34,A35 | B14 | P1,P2 |
| CC4 | A56 | B87 | |

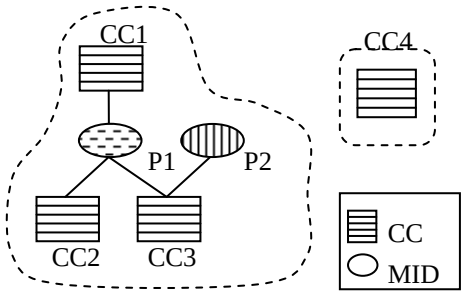


Figure 3
Building cookie networks

This section provides the detailed structure of a system processing weblogs. Figure 4 contains all the mayor components and data files used. The role of the different components is summarized in this section, while data file are described in Section 4.

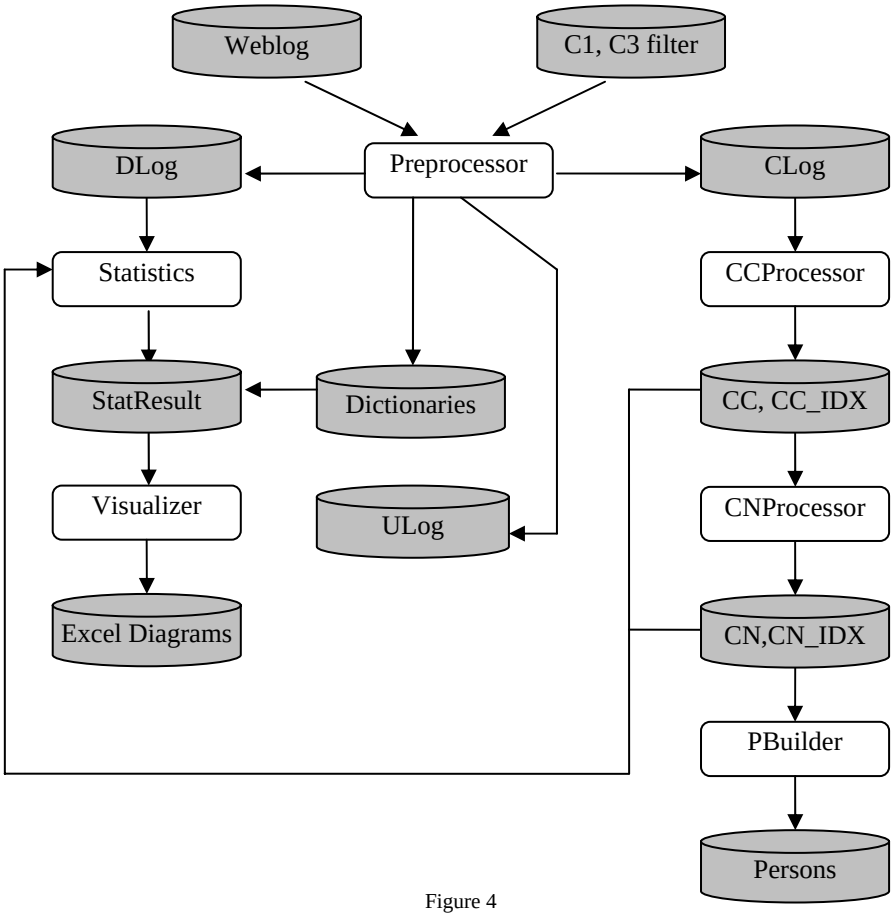


Figure 4
Weblog Processing

The processing components are the following:

- Preprocessor
  - *Filter:* Due to various system errors weblogs can contain invalid entries which have to be filtered.
  - *Compressor:* Due to the large size of weblogs, it is very important to compress them and select the most important information for further processing.
- CC Processor
  - *CC Builder:* CC-s have to be built in memory which causes that only a limited span of CC-s can be managed at a single time.
  - *CC Merger:* in order to increase the span of CC-s, CC Merger has been introduced to merge the previously estimated and stored CC-s with the newly built one.
- CN Processor
  - *CN Builder and Merger:* it resembles the CC Builder and Merger, however it uses CN-s instead of CC-s and build and merges CN-s in one step.
- Real person identifier (PBuilder): it finds the real persons belonging to the CN-s using several business rules. The most important rules are the following:
    - MID-s belonging to different CN-s belong to different real persons.
    - A real person cannot browse the web from different computers (having different IP-s) in a short time span.
    - If two MID-s can be found in three different CC-s, they belong to the same real person.

Besides the data processing components, there are several others belonging to the system. One is the statistical component which can create several statistics based on CC-s and CN-s. The Visualizer component can be used to build Excel diagrams automatically from statistical results.

The following section describes the data formats used by the different components.

# 4 Structure of Used Data Files

Section 2 described the contents of a weblog file having 392 byte-sized records. Since web auditing produces 10-20 GByte weblogs (containing 40-50.000.000 records) daily, it is a key factor to compress and convert the contents of the logs efficiently. The domain for the different elements of weblogs can vary; Figure 5

compares how many different C3 cookies, MID-s, Ucodes and Web clients can be found in weblogs. It is important to note that the web client contains not only the browser's name but also gives information about the client's operating system and the browser plug-ins installed. In fact the amount of different C3 cookies, MID-s, Ucodes and Web clients and is the size of the appropriate dictionaries. Figure 6 compares the amount of MID-s found in weblogs to the size of the dictionary, it is important to note that in most cases when a MID entry is processed during the compression, in most cases it can already be found in the dictionary according to Figure 6.

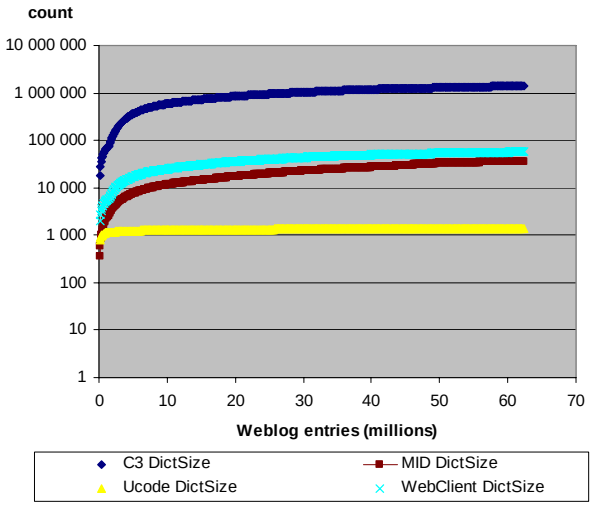


Figure 5
Increase of the domain of weblog elements

In order to build CC-s and CN-s, the C1, C3 and MID values are needed, so the Clog file contains these values. C1 cookies can be converted into two different 4 byte-sized numbers, however C3 cookies and MID-s can only be compressed with dictionaries because their internal structure is currently not available. A dictionary contains key-value pairs (uncompressed and compressed data). The size of the uncompressed C3 cookies is 22 bytes according to Figure 1, while the compressed C3 cookie is only 4 bytes. 4 bytes are enough, since $2^{32}$ (around 4.000 million) different values can be represented on them. It means that if there would be 3 million Internet users in Hungary, each of them could have 1200 different C3 cookies in 2 month time (C3 cookies are stored in the dictionaries for 2 month, if a cookie is not used for more than 2 month, it is removed from the dictionary). Those data not required to build CC-s and CN-s, but needed for the statistical component will be stored in a separate file called as the Detailed Log (DLog) file. These data are the following: IP, Time, Ucode, Web client. Ucode and Web client have to be compressed with dictionaries. Url values are stored unchanged in the

ULog files. The size of the most important files – without dictionaries is the following:

- CLog: 16 byte
- DLog: 16 byte
- ULog: 100 byte

According to our measurements, a 1.9 gigabytes sized weblog file could be compressed without losing important information to data files having 200 megabyte total size (10 percent of their original size).

At the end of the system architecture Section 5 will describe the software architecture of the system.

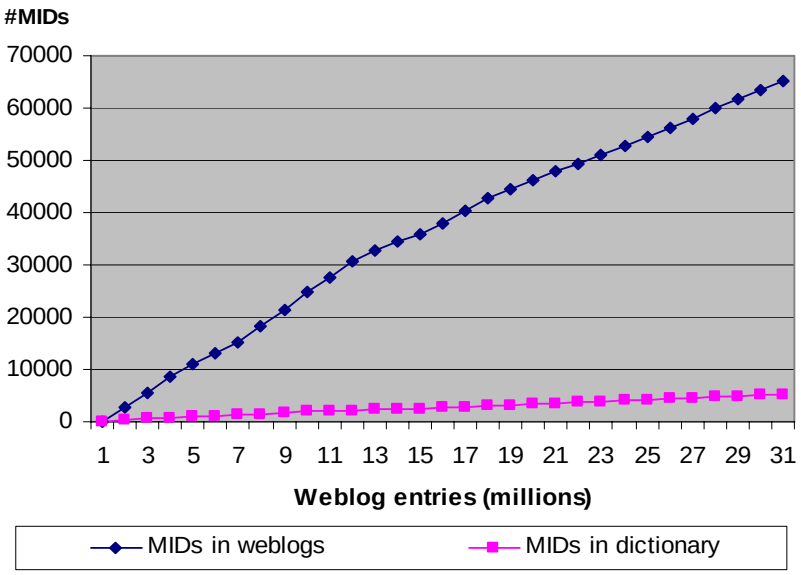


Figure 6
Decrease of domain rise

# 5 Software Architecture

Section 3 contained the description of the major components of the system, while Section 4 detailed the data files used along with the possible data compressions and conversions. This section describes the layers of the system (as depicted on Figure 7).

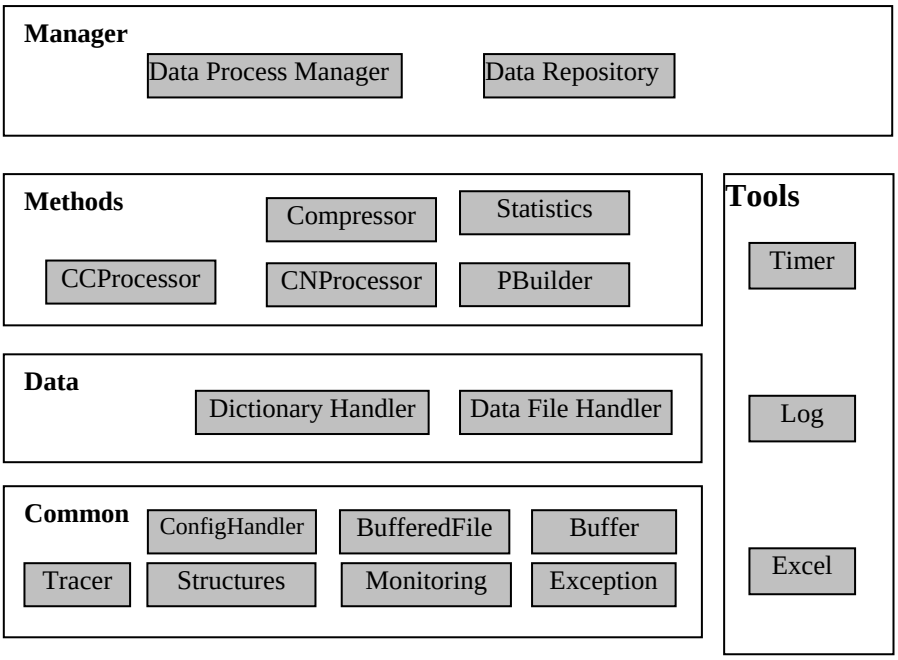


Figure 7
Software architecture

The Common layer contains those modules that have basic functionalities. It provides modules for efficient buffer-based reading and writing of data files used. Reading is available both in synchronous and asynchronous mode, while writing is supported in synchronous mode. This layer contains another module for tracing function calls with their parameters in an easy-to-read form. Another important functionality of the Common layer is monitoring of the state of the system in order to be able to restart or continue weblog processing based on system states. Each module of the system (such as CCBuild, Compress) can have its own configuration file, so a ConfigHandler module is needed to manage the configuration files.

The Data layer contains the different reader and writer functions for the supported data files (such as weblogs, CLogs, DLogs, ULogs and dictionaries). An efficient search algorithm has been implemented on the dictionaries.

The Methods layer is responsible for the processing of weblogs (compressing, building cookie chains and networks, finding real persons, building statistics).

The last layer, the Manager contains the data repository to locate the appropriate physical files and give them to the processing modules of the Methods layer. It is responsible for the scheduling of the weblog processing.

The Tools layer contains secondary components that do not belong to the final system, however they can be used to obtain measurement results. It contains functions to display Excel graphs, has a timer and a measurement logger.

**Summary**

This article provided a detailed description about the architecture of our system to process large weblogs in onder to separate the activities belonging the separate real-life persons. By assigning Internet users to real persons it is possible to estimate the number of visitors of web sites which is very important for both the sites and their advertisers. In order to find the real persons complex business logic has to be used which requires an efficient compression of the weblogs. Our measurements proved that the weblogs can be efficiently compressed without losing important information to about 10% of their original size.
Future plans include building cookie networks differentially, optimizing the cookie chain and network file formats and the estimation of real persons belonging to cookie networks.

**Acknowledgement**

This work was accomplished with active cooperation of Median Public Opinion and Market Research Institute and supported by the Mobile Innovation Center, Hungary. Their help is kindly acknowledged.

**References**

[1] Median webaudit, www.median.hu

[2] Tárki auditonline, http://www.auditonline.hu

[3] Szonda Ipsos audit, http://www.szondaipsos.hu/hu/ipsos/gemius

[4] Coremetrics auditing, http://www.coremetrics.com/technology/first_party.html

[5] CMP United Business Media, Privacy Statement, http://www.cmp.com/delivery/privacy.html

[6] Le Beaumont Language Center, Privacy Statement, http://www.alihk.net/beaumont/clubs/privacy.htm

[7] Siteframe 5.x audit logging, http://siteframe.org/p/logging

[8] Csaba Legány, Attila Babos, Sándor Juhász: Cookie-Chain-based Discovery of Relation between Internet Users and Real Persons, 5th International Conference on Information System Development, (ISD 2006)

[9] Jack Powers: Counting Clicks: Auditing Your Web Site Activity, Fall 96 Internet World, New York, December 12, 1996 http://in3.org/course/clicks.pdf

[10] AboutCookies.org, a guide to deleting and controlling cookies, www.aboutcookies.org

[11] Cookies: The Perfect User Identification Snack, http://www.clickstreamdatawarehousing.com/article06.html