_____

Possibilities of visualisation geolocation of web pages visitors

Lukáš Svoboda

Katedra geoinformatiky, Přírodovědecká fakulta, Univerzita Palackého,
tř. Svobody 26, 771 46, Olomouc, Česká republika
l_svoboda@centrum.cz

Abstrakt. Příspěvek se zabývá možnostmi vizualizace informací o geografické lokalizaci návštěvníka webové stránky. Stručně je popsána problematika monitoringu návštěvnosti z hlediska používaných technologií a následného využití takto získaných informací. Představeny jsou metody zjišťování geografické lokalizace návštěvníka, je zhodnocena jejich přesnost a nastíněno praktické využití. Na základě testování vybraných českých i zahraničních služeb pro monitoring návštěvnosti webového serveru byly zjišťovány možnosti, použité metody a technologie vizualizace geografické lokalizace návštěvníka. V závěru jsou nastíněny další možnosti vizualizace, jejich přínos pro uživatele a očekávané trendy v oblasti geografické lokalizace návštěvníků.

Klíčová slova: geografická lokalizace, geolokace, monitoring návštěvnosti, vizualizace

Abstract. The paper tries to introduce the potentialities of a visualisation geolocation of the web pages visitors. The problem of the web analysis technology, used metrics and the subsequent purposes is briefly described. There are introduced the methods for an identification the geographical location of the websites visitors, there is evaluated an accuracy and a reliability of this methods and a practical usage is described. Based on the testing, the number of the selected web analysis services was compared by the possibilities, used methods and visualisation technologies of websites visitor's geographical location. At the end, there are designed the new possibilities of a visualisation for the websites visitors geolocation services and there are also sketched the expected trends in the geolocation technologies.

Keywords: geographic localization, geolocation, web analytics, visualisation

1   Introduction

Owners and webmasters need more information such as who is visiting their web site, which web pages they are visiting or which pages they are ignoring. The answers to these questions are available through a technology called web analytics. One of these questions could be: From where are visitors of the web page? The key to answer this question have Internet geolocation technologies. Information about the physical location of web visitor can be used for wide variety of purposes, including targeted marketing, restricting digital content or localized tools for Internet users. Important question is how accurate and reliable these technologies are.

2   Web analytics

2.1   Data Collection Methods

Currently, web analytics services use two of the leading methods to gain information about your web site. The first method involves using a web server log file, which contains some basic information about the activity on web site. The second method involves collecting data from the visitor's machine by using client-side tagging to create a more detailed and customizable kind of log file than the standard logging available from server software. Many web analytics services use both web server logs and client-side tagging. Web server logs can help them to obtain IT-based metrics such as spiders, downloads, bandwidth, and errors. Client-side tagging can help them to get business metrics such as screen resolution and java enabled browsers. There are many other differences in the data collected by these two methods that may or may not be relevant to analytics needs [8].

In general, the benefit of web server log files is that visits of all website parts are collected. Log files capture all downloads and non-HTML files in addition to HTML files. One of the drawbacks is fact, that log files cannot record repeat requests when a page is accessed from a caching or proxy server.

_____

Client-side tag (script) runs each time the page loads, the benefits are accurate visit and page counts, even when pages are loaded from a caching or proxy server. Problem of client-side tags is that the web browser must be enabled to run the scripts. Client-side tags also require time or software to embed the script in each tracked page.

2.2   Used metrics and goals of web analytics

Most of the web analytics providers use the metrics recommended by International Federation of Audit Bureaux of Circulations (IFABC), which is an international voluntary cooperative federation of industry-sponsored organizations established to verify and report facts about the circulations of publications and related data. The basic metrics for measurement of website traffic are [1]:

- *Page impression* - a file, or a combination of files, sent to a user as a result of that user's request being received by the server.
- *Unique user* - the total number of unique combinations of an IP address plus a further identifier. Sites may use User Agent, Cookie and/or Registration ID.
- *Visit* - a series of one or more page impressions, served to a valid user, which ends when there is a gap of 30 minutes or more between successive page impressions for that user.
- *Session* - a series of page impressions served in an unbroken sequence from within the site to the same user.

Visit duration - the total time in seconds for all visits of two or more page impressions, divided by the total number of visits of two or more page impressions.

Web analytics are the set of tools for realization of the web site goals. Every web site has primary goals. Some sites sell products or provide information. Others specialize in user videos to provide an entertainment. For example, the main goal of a self-service web site can be to save costs related with other methods of customer service (such as email and technical support) versus generating revenue. Many sites serve a combination of purposes. For example, a theater's web site typically provides information about the programme and sells tickets to its plays. Common goals of web analytics are improving visitor retention, decrease costs or making web site more valuable and comfortable to visitors.

3   Geolocation of web visitors

Geographical location of Internet user is common tool of the web analytics services, regardless of price these solutions. Owners and webmasters have a lot of purposes to know information about the physical location of a web site visitor. Some of these purposes are introduced in next list:

- *Restricting digital content, different jurisdiction, criminal investigations* – obeying licensing regulations of the web content in different countries (Internet gambling, video of sporting events, pornography). By identifying where online visitors really are, geolocation can protect companies from participating with terrorist organizations or trading with banned nations.
- *Security and network efficiency* – network optimalization, access to content on the nearest node of network
- *Click fraud detection* – detecting possible credit card fraud by comparing the user's location to the billing address or the shipping address provided
- *Target content* – web sites can show different web content based on your geolocation or other information. It could be used for censorship as well as web site's language versions.
- *Geo marketing (targeted advertising)* – companies may provide web content or products depending on user location. Advertisements and content on a website may be tailored to provide the information that a certain user wants.
- *Spam fighting* – ISPs can identify or flag messages and posts from traditional spam countries or spam centers.

3.1   Geolocation techniques

Because every host is directly connected to the Internet is identified by a unique IP address, the most of geolocation techniques are based on linking geographic locations with IP addresses. The IP

_____

geolocation techniques can be classified into three avenues of approach [4]. The first group of techniques uses databases for look-up of IP address or domain name. The second approach tries to identify and correctly interpret any geographic codes present in a domain name. The last group is based on routing and timing information in network. Many of the techniques are in use commercially, often in combination. Exactly which techniques can be used depend upon context since, in different situations, different information is available.

Information about an IP address is most easily obtained by look-up in public whois databases, which contains connections between logical Internet identifiers (IP addresses, autonomous system numbers and domain names) and real-world entities. This allows determining the entity to which an IP address (AS number, domain name) is registered (e.g., an ISP, company, organization, end user, etc.). Contact information usually includes email address, telephone number and mailing address that can be used for geographic location.

Generally techniques based on whois databases have some drawbacks. The main problem is that not all target hosts are located at or near the address of the organization to which the host's IP address is registered. In addition, the whois database data can contain incorrect or outdated records.

Another way to define the geographic location is interpreting any geographic codes from domain name, application settings or directly from user. Domain name can contain a definition of country (e.g. http://www.domain.cz is probably located in Czech Republic) and/or another geographic location, but the usage is limited by researchers geographic skills (http://www.olomouc.cz, http://www.ci.ontario.ca.us, etc.).

Some information (language, charset, operating system version) can be found in HTTP headers that are sending by browser. Similar headers are send by email client. Location data can be also entered directly by user on the website (requests for a ship-to-address, local weather forecast, list of countries on web entry page) which enabled associate it with a client IP address. These techniques have also some drawbacks. HTTP headers can be filtered by privacy tools (Tor, Privoxy, etc.). Any application or user submitted data may not be correct and reliable [4].

Third group of techniques is based on the premise that the delay of packets traveling between a pair of network hosts is function of the geographic separation between the hosts [5]. The command line tool ping is used to send an echo message to a host and time until received can be measured (this technique is also know as GeoPing). If it is difficult to find the location of target host, it may be useful to locate the closest hosts by tools for monitoring of routing packets. The similar method to locate user is recognition of ISPs network topology (geographic layout) and internal routing rules using information from ISPs. Limitation of these methods is fact that some target hosts not respond to echo messages or messages for tracing.

New geolocation technique use Java applet for determines user's IP address even if is user protected by privacy tools (Privoxy, Tor, etc.) and connect by anonymous public proxy servers. Java applet makes a non-proxied connection back to the web server and extracts real IP address of user.

4   Possibilities of visualisation geolocation of web pages visitors

4.1   Present state

Geolocation tool is a common component of the web analytics services that use various techniques for measuring and visualisation. The current techniques for geolocation visualisation can be classified into four groups:
- *table*
- *chart*
- *raster map*
- *dynamic map*

Tables or charts are the most popular techniques for geolocation visualisation. Is fact that these methods have simple requirements for processing and visualisation. The tables can be sorted by the visitors count (total count or percentage) or by the geographic names (countries, regions, cities). National flags are often used for better orientation. As a common addition are used the charts (pie or column type). It is also everything what majority of the web analytics services can offer.

The map tools in web analytics can offer more transparent and varied content than the tables or charts. Identification of visitor's location and monitoring geographic trends in a map (especially dynamic map) can be easier and more transparent for a web analytics user. Identified trends can be used for targeted advertising, extending language options on the website and others.

_____

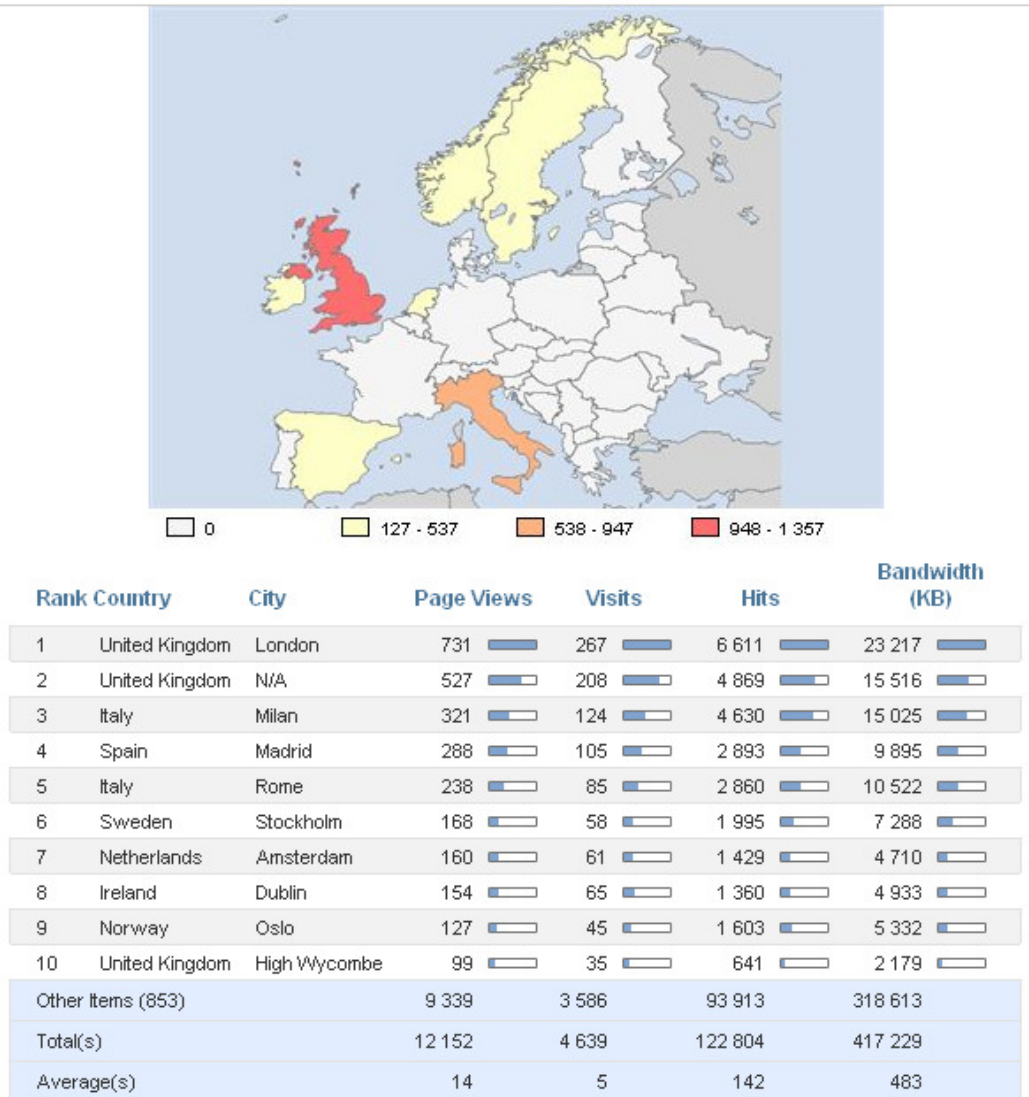**Table 1.** Summary of selected web analytics services

Service	Price	Technique	Visualisation geolocation	Geolocation scale
AWstats	free	log	table	country
ClickTracks	paid [1]	log	map, table	country, region
Gemius SA	paid	script	map, table	country, region, city
Google Analytics	free	script	map, table, chart	country, city
Navrcholu.cz	free, paid	script	table, chart	country, city [2]
NetAgent	free	script	table, chart	country
TopList.cz	free	script	table	country
SiteCatalyst	paid	log	?	country, city
SmarterStats	free, paid	log	map, table	country, city
TopList.cz	free	script	table	country
WebTrends	paid	log	table	country, city region

[1] geolocation is possible only in paid version (Optimizer, Professional, Pro+JDC)
[2] only in Platinum version

In the Table 1 you can see the list of Czech and world most popular web analytics services. Only four services of them use maps for presentation of geolocation web site visitors. Two of these (ClickTracks, SmarterStats) display map in the form of cartogram without any dynamic features. Both don't use the labels of the states or cities on the maps and even in the case of Clicktracks map there are no symbols of cities on the map called "European cities" (Fig. 1).

European Cities

Rank	Country	City	Page Views	Visits	Hits	Bandwidth (KB)
1	United Kingdom	London	731	267	6 611	23 217
2	United Kingdom	N/A	527	208	4 869	15 516
3	Italy	Milan	321	124	4 630	15 025
4	Spain	Madrid	288	105	2 893	9 895
5	Italy	Rome	238	85	2 860	10 522
6	Sweden	Stockholm	168	58	1 995	7 288
7	Netherlands	Amsterdam	160	61	1 429	4 710
8	Ireland	Dublin	154	65	1 360	4 933
9	Norway	Oslo	127	45	1 603	5 332
10	United Kingdom	High Wycombe	99	35	641	2 179
Other Items (853)			9 339	3 586	93 913	318 613
Total(s)			12 152	4 639	122 804	417 229
Average(s)			14	5	142	483

**Fig. 1.** Visualisation of visits, Europe (ClickTracks)

More interactivity is provided by Gemius SA and Google Analytics services. Gemius SA service can display the dynamic labels on cartograms. When the mouse is over the information about name of a city (region) and the number of visitors is displayed. The web analytics reports are filtered after the selection of region or city in the map. Google Analytics provides similar possibilities and adds continuous switching between different countries and cities in the dynamic map. The cartodiagram form is used for displaying of cities (Fig. 2).

In all of the discussed web analytics map visualisations is missing the scale from cartographic point of view. The others common cartographic mistakes are missing labels, bad color graduation or incomplete legend.
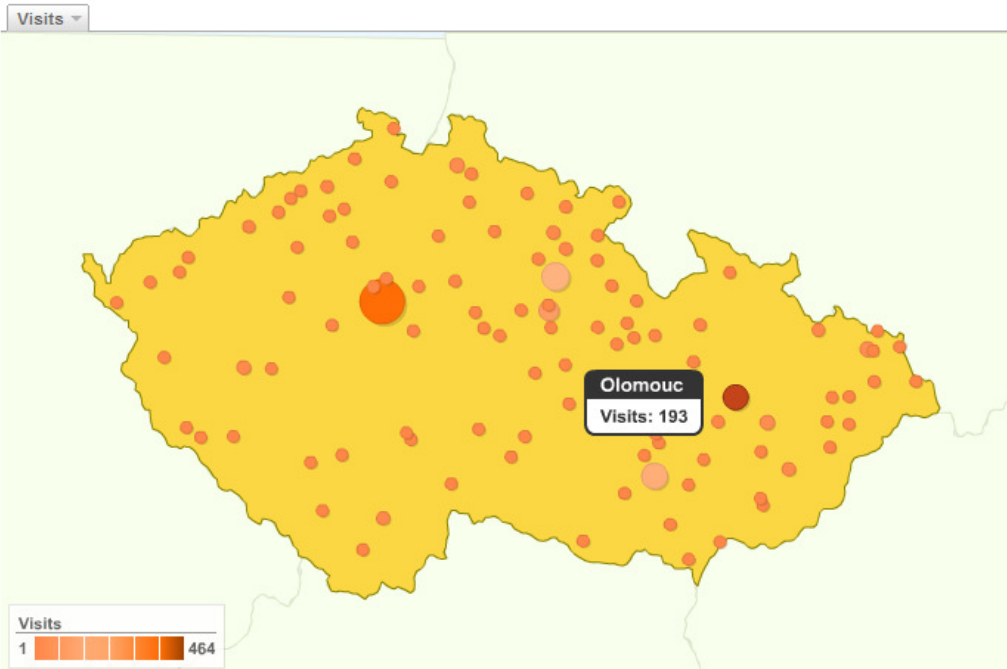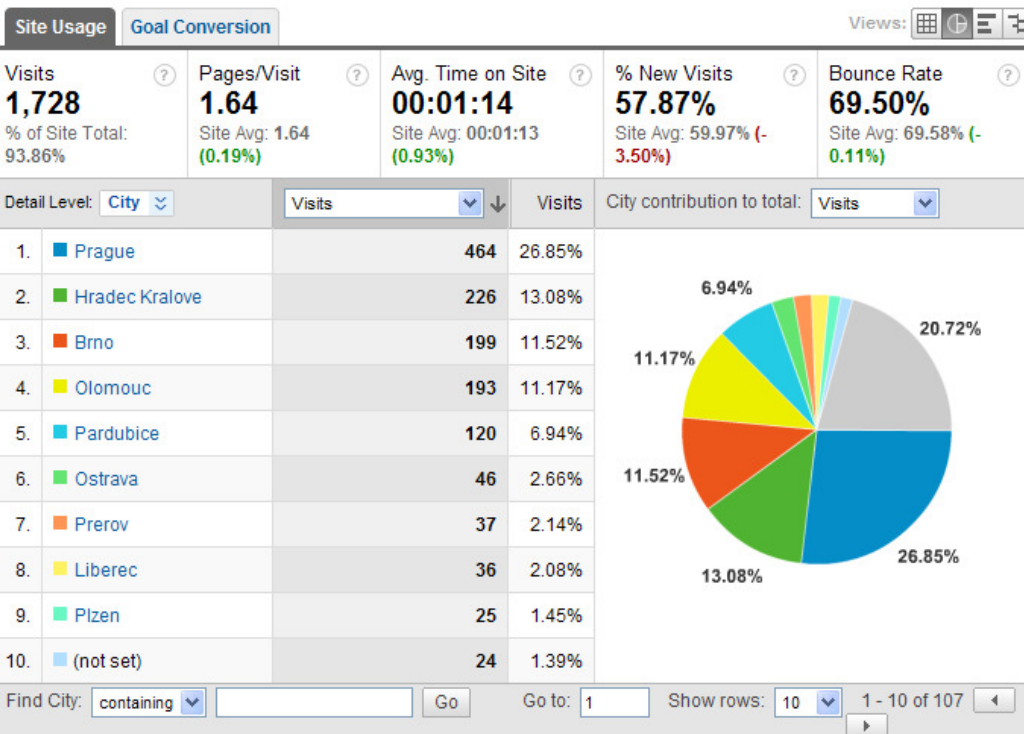
_____

Fig. 2. Visualisation geolocation of the website (Google Analytics)

4.2 Evaluation of possibilities

Generally, web analytics services don't use all possibilities for visualisation geolocation, although current web technologies provide a wide variety tools for visualisation of dynamic features (for example Adobe Flash, Adobe AIR, Microsoft Silverlight). Google Analytics is the best web analytics tool (from the services mentioned above) from geovisualisation point of view, but it also has some space for the improvements. An implementation of Google Maps could be logical improvement of this service. Google Maps could bring stronger map tools, real geographical background layers or selecting web analytics reports by non-standard geographical regions. Interesting fact in this context is that advertising tool Google AdWords has Google Maps integrated for the geographic targeting campaigns (Fig. 3).
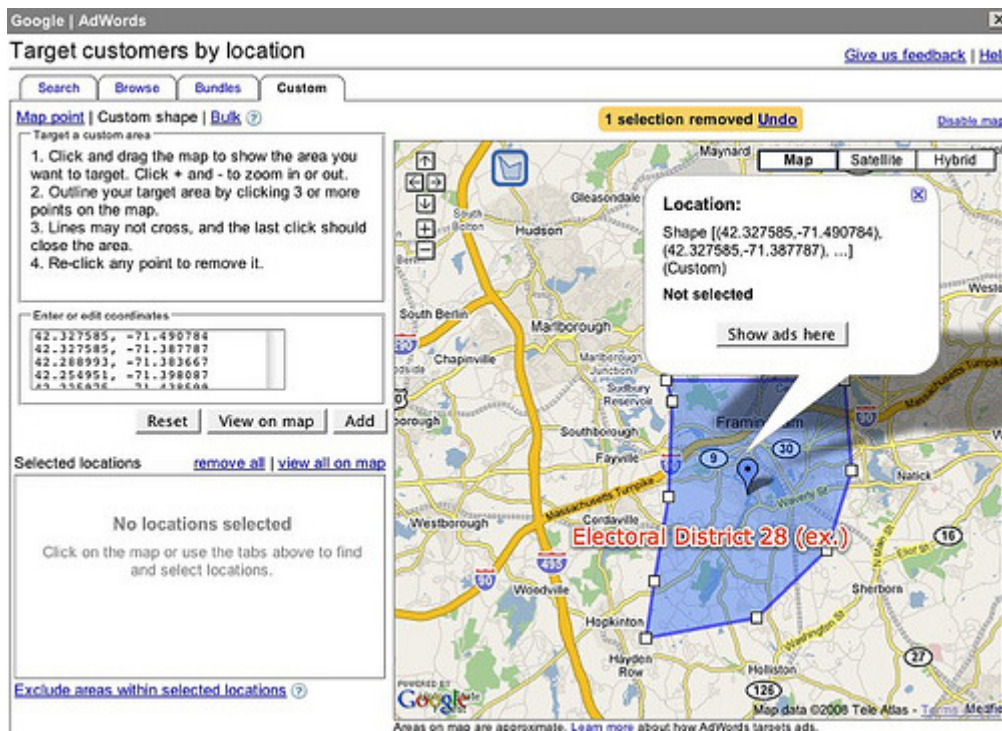

Fig. 3. Advertising targeting by geographic location (Google AdWords)

5 Conclusion

The web analytics are fundamental information for web site administrators and owners that can be used for a wide variety of purposes. An irreplaceable part of web analytics is information about geographic location of web site visitor. Currently, the most widely used geolocation technology is based on the translation of IP addresses into geographical locations, by the use of information stored by the provider of the geolocation service (whois databases). Other techniques include interpreting geographic codes in domain names or HTTP headers, network routing methods and HTTP refresh timing. All the techniques have some limitations and drawbacks and they are in use often in a combination. Exactly which techniques can be used depend on context, in different situations different information is available.

References

1. International Federation of Audit Bureaux of Circulations. *Web Measurement Standards* http://www.ifabc.org/standards.htm

_____

2. Jašek Pavel. *Webová analytika v internetovém obchodování.* Vysoká škola ekonomická v Praze, 2008, Praha

3. Lašek Ivo. *Audit návštěvnosti webového serveru.* České vysoké učení technické v Praze, 2007, Praha

4. Muir James A., van Oorschot, P. C.. *Internet Geolocation and Evasion*. http://cs.smu.ca/~jamuir/papers/TR-06-05.pdf

5. Padmanabhan Venkata N., Subramanian Lakshminarayanan. *An Investigation of Geographic Mapping Techniques for Internet Hosts*. http:// research.microsoft.com/~padmanab/papers/sigcomm2001.pdf

6. Peterson, Eric T.. *The Future of Web Analytics*. http://www.webanalyticsdemystified.com/sample/Web_Analytics_Demystified-_The_Future_of_Web_Analytics.pdf

7. Svantesson, Dan. *Geo-location Technologies: A Brief Overview*. http://svantesson.org/svantesson20040906.doc

8. WebTrends. Guide to Web Analytics http://product.webtrends.com/WRC/7.1/Documents/WebAnalytics.pdf