Identification in an online learning environment

Z. BALOGH
Research assistant, Corvinus University of Budapest
zoltan.balogh.jr@gmail.com

ABSTRACT

The students of the Corvinus University of Budapest are using the University's e-learning system almost every day. Probably most of them are not aware how much information they share with the website. In my research I am trying to find the answer why the gathered data is important for the companies and can be converted to valuable information. In this paper I am describing the first phase of my research and the preliminary results.

Prologue

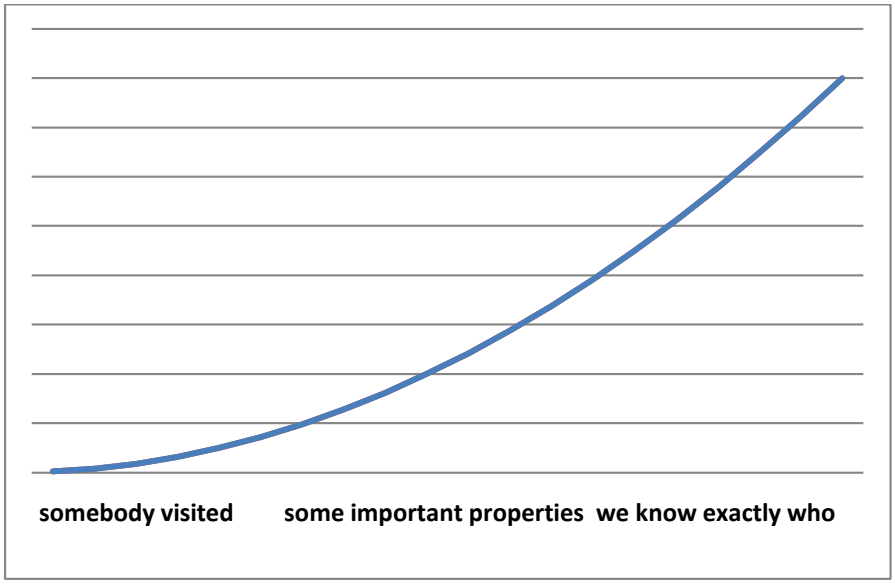
Tim Berners-Lee created the very first website in '91, in 2013. march 18. he was awarded the Queen Elizabeth Prize for Engineering, because „his innovation has revolutionised the way we communicate" (1) Probably he did not think how popular the Internet and the HTML will be 20 years later.

According to the statistics in 2012 more than one third of the entire human population were using the Internet (2) and by the year 2015 the worldwide Internet traffic will approach the astonishing 1 zettabyte per year. Just for comparison it is all the digitally stored information as of 2010. (3)

Problem statement

The Internet enabled the development of whole new industries and also helps existing businesses to expand. Marketing was one of the winners of the expansion, because with the Internet it has never been easier to send targeted advertisements to customers. With profiling the visitors of web sites it is possible to categorise them by their common properties or interests. But how does profiling works. How many information do we need of the visitors to categorise them properly?

On the chart of identification (1. chart) ranging from somebody has visited the website to we know exactly who is the current visitor, there are different levels of identification.



1. chart, amount of acquired information of the visitor (self-edited)

The goal defines the level of identification. For the business, it is not necessarily important to know the exact identity of the visitor it is enough to know his essential properties and preferences.

Research issue

In my research I have developed a data collector software, that I integrated into the e-learning system of Corvinus University of Budapest, that collected all the available data of the visitor students' software and hardware environment. It also saved their geographical position and available data on Facebook with their prior consent.

The e-learning system (aka Corvinus Moodle) has 8542 users in the spring of 2012 and the total number of students of the University is 17 869 and the number of teachers are 867, which means that approximately 45% of the university teachers and students are using Moodle.

The data collection started on 18. april 2012 and lasted for 1 month. The application gathered more than 840,000 records. (1 record means 1 page download)

Device-browser

Before going on, we need to define the term of device-browser. While browsing the web websites can store data on the client side. A popular way of storing this data is the cookies (which is basically a text file, that can be read by the website that previously had saved it). Unfortunately a browser cannot read the cookie of other browsers, which means, that if the same person from the same device changes browser, in my database he will be treated as a different individual. I use the term „device-browser" for a browser on a device of the visitors.

In the current state of my research, my aim is to find a way how to link different device-browsers. In this paper I am expounding how I started to link the different-device browsers, what are the results and what are my future plans.

Why is it important?

In 2007 Facebook introduced the „Beacon" marketing program, that would help users keep their friends better informed about their interests and would help drive more sales to the participating sites. But this application showed in status updates and with photos of what items their friends had bought or reviewed. That's why users complained, that friends could learn of holiday gifts they had bought at the online retailer. (4)

In 2010 Facebook introduced the "Like buttons" and other widgets that can be implemented into websites. These modules can also send some information to the servers of Facebook about the visitors.

There is still a vivid battle for the attention of the visitors. Years later, many of the top influencer companies of the IT industry are decided to take up the gauntlet and join the battle, including Microsoft with So.cl, co-founders of Twitter and Blogger with Medium. (5)

These companies are fighting for the customers, because a well-functioning social site can collect valuable information on its visitors. The information of the visitors can be categorised and targeted advertisement can be sent to them.

Also there are other purposes for identification of the visitors:

- Fine tuning online services / Improve user experience : if the preference of the user is known, then a higher level of user experience can be given to him
- Targeted advertisements : if the interest of the visitor is known, then more relevant information on products can be sent to him by e-mails
- Finding somebody on the Internet : people can be found in the real world with the help of the above mentioned techniques

The different purposes claim different levels of identification. In the first two cases it is not necessary to know the exact identity of the visitor. In my research, I am dealing with the data collection part and I used ideas from Bursts – The future is predictable. The book says that most people behave in a predictable way. (6)

Approaches of identification

According to the literature I found, that there are two different approaches for identification of individuals:

- Top-down: Starts from the entire population and drills down to the individuals (7)
- Unique in the crowd: The theory is based on the uniqueness of human mobility (or behaviour) (8)

Since there is no behavioural data for all of the visitors of Corvinus University, the top-down approach cannot be used here, so I decided to go with the unique in the crowd approach. It is possible to identify the individuals by the features of visitors' software and hardware environment used for browsing visitors' online behaviour

Methods of identification

During my research I found many identification methods, which can be categorised in the following way:

- Technological
- Account based
- Social site linkage or OpenID usage
- Location based
- Software & hardware environment based
- Browsers' cache mechanism based
- Based on behaviour

The account based or social site linkage identification is not in the scope of my paper, because in case of a unique id (like e-mail or Facebook account) the identity of the visitor is obvious.

Account based

The visitors are identified by their e-mail address or their self-chosen unique login name. If the user is signed in, it is pretty easy to detect what he is doing on the website or get in touch with him.

Social site linkage or OpenID

Today it is popular to use Facebook as the authenticator for websites. With the installed Facebook login application it is possible to register and authenticate visitors. It is also much more convenient for users, because they don't have to waste their time with filling the websites' own registration form, not even mentioning they don't have to memorise an additional password. The registration is just a couple of clicks, the visitor can grant access for the website to fetch his required data from Facebook and authorise the user.

As I mentioned before this is a pretty convenient for the customers, but this has also disadvantage for the visitors and that's why some people are reluctant to use it. The visitor has to restrict his Facebook profile from other people to check it out. And sometimes it is not enough, by any chance the visitor is the friend of the owner of the website, then he can get sensible information on him.

Location based

It is known that the IP address can be resolved to a rough location. The ability of changing places of people has its limitations and they usually use the internet only from a few places.

If the user approves the website to use his geolocation info, the location will be much more precise. (We can use the IP address, the ISP and the location of the user to identify in his session)

Software and hardware environment based

With the help of Javascript and Flash technologies, it is possible to get a thorough feature list of the visitors' hardware and software components, from which a fingerprint can be generated that helps to identify the visitors browser-device.

Browsers' cache mechanism based

Cache is a great help for the browsers, with which the requested webpage can be shown much faster and bandwidth can be saved. During the browsing all the resources (including images, scripts, CSSs and HTMLs) are stored temporarily in the hard drive of your computer. If the visitor wants to visit a webpage that he has visited before, then the browser will check its cache first. If it finds the requested resource there, then obviously it will not download it again, thus saving bandwidth. But there is a mechanism that tells the browser if there is a newer version of the resource on the server that needs to be downloaded. This mechanism can be used to identify visitors. (9)

Based on behaviour

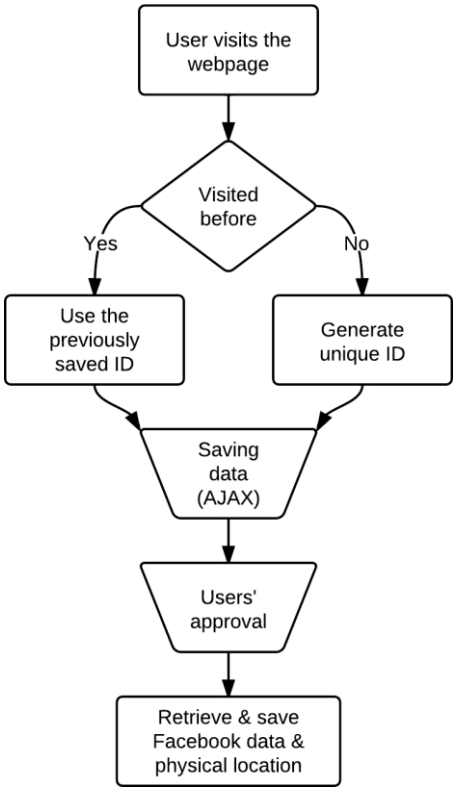
This method includes the identification of the visitors by their behaviour, which means that everything you do in the online world can define you.

If we gather lots of data of the visitors, it is possible to find out, what they are interested in. Profiling is a method that puts the similar individuals into groups of the same property.

Research & result

In April 2012 I implemented a visitor tracker application into the e-learning system of the Corvinus University of Budapest with the approval of University. The application saved all the available attributes of the visitors' software and hardware environment and their geolocation and public data on Facebook with their consent. There is an entry in the database for every click of the visitor. When the visitor downloads the very first page, he gets a unique ID that is stored at the client-side in a cookie and in localStorage[1].

[1] In 99,36% of the records supports localStorage in the sample



After every page download the application examines the content of the cookies and the localStorage and if it finds valid existing tracking data, then it means that the user has already been visited the webpage from the current browser in the current device.

Browser do not share cookies, which means if the visitor downloads a page with Internet Explorer and later does the same thing with Chrome from exactly the same machine the application is not able to track him. The other browser is not able to read the browser's cookie or localStorage, but the hardware and the software environment is still the same. (Though might not be detectable with an older browser)

If the visitor downloads the page for the first time, the application generates a unique ID and saves it on his browser, so later this can be used to track him.

Behaviour based identification

The content consumption from different device-browsers might be significantly different. But from this point of view the e-learning system of the Corvinus University is special, because the visitors can get to the desired content through the same pages (nodes), the chance of direct linking is low. So we can say that visitors browsing from different device-browsers can have similar browsing path or history. For example the students can use the e-learning system for learning at home from their laptops, but they can also check the test results with their mobile phone. In this case they have to login, then click on the site of the course and then they can click on the results. In this case there is no chance of direct linking, because the students don't know the link of the results prior to publish.

This means that it is possible to build the path tree of the visited pages of the visitor and can search the database for similar patterns.

In real life test the following points should be considered:

- The tree path of the visited webpages
- The order of the visited webpages
- The used device-browser for visiting the webpage
- The process of identification

As mentioned before the visitors can be identified by the software and hardware environment of the used device, by the behaviour of the visitor and the combination of these.

As a first step select a random record from the dataset, then draw the tree path of the visited pages. Then search the database for similar patterns, only those visitors were selected, whose tree path had at least 2, 3, 4 or 5 visited pages. If the number of the matching visited pages is low, then the algorithm results lots of similar visitors that are probably classmates or alternatives device-browsers of the main visitor. If the number of the matching visited pages is high, then potential device-browsers might be skipped. That's why I decided to keep this number low.

After having the group of the potential classmates and device-browsers, started to compare them with the sessions of the main visitor according to the following rules:

- Time dimension
    - short time frame: in a short time frame, most of the stored data can be used for identification (it is unlikely for a visitor to change IP address, screen or OS within few minutes)
    - long time frame: in a long time frame, none of the stored data can be considered permanent, so it cannot be used for identification (for example: during the period of the data collection most of the visitors upgraded their browser, but it is also possible that somebody changed his screen or laptop)
- IP address
    - Non-Corvinus University IP address
        - Fix: can be used for identification
        - Dynamic: Can only be used for identification in short time frame. Most people are sharing the internet at home, with a SOHO router. This means that it is using one public IP address, that's why in short time frame if two different device-browsers are accessing the internet, then it is the same person or somebody that lives in the same place
    - From the domain of Corvinus (all the IP addresses of Corvinus starts with 146.110)
        - Fix: Most of the IP addresses at the Corvinus are fix. The fix IP addresses can be given to employees, campus building classrooms, students in the dorm.
        - Dynamic: Smaller part of the used IP addresses is dynamic. Dynamic addresses can be given to campus wifi access and VPN. Assuming that the users are not sharing the given IP addresses, these can only be used during the session, because the users get a new address, each time they connect to the network.

- data of the ISP
    - the name of the ISP for the connection can be resolved from the IP address can also be used for identification (it is not possible that a user has many internet connections at home)
- geological position
    - position from IP address: there are online databases with which it is possible to resolve IP addresses into geographical locations. The free version support city level resolution, so it can be used in case of a visitor who is not from Budapest.
    - geolocation
        - in case of wired connection: in case of wired connection the geolocation position is superficial, it only shows city level resolution
        - wireless connection or equipped with positioning capable device (A-GPS, GPS, GLONASS, Wifi)
- font types: the font types installed into the device used for browsing. The fonts are browser independent, that's why it can be used to identify different browsers on the device
    - device used for browsing: in case of tablets and mobile phones the hardware environment can be the link between different browsers
- visited web pages: the common web pages in the visitors' tree path can be a starting point to connect the device-browsers
- In some of the points I used the specificity of the network of the Corvinus University.

I tested the above mentioned rules on a group of 24 individuals. I used the combination of the above mentioned rules and in 25% of the cases I could link the device-browser with full certainty. In 8% of the cases I found the probable device-browser.

Conclusions

An average of 3,93 kB of data is saved with every click of the visitor in my database, and there is an average of 19,89 records of every browser-device in my database. So there is 80 kB of data of every device-browser and there is 300,5 kB of data of every user in my database.

The result of the query shows, that with browsing visitors are sharing lots of information about themselves and about the device used for browsing, and other researches claim that it can be used to conclude aspects of their personality, including sexual behaviour or intelligence. (10)

By browsing the web, we are offering the sensitive points of our personality and habits to the social sites and the visited websites. The more we use the web, the more accurate profile they will have on us.

Cited resources

1. **The Royal Academy of Engineering.** Queen Elizabeth Prize for Engineering. *Announcement of the first winner of the Queen Elizabeth Prize of Engineering, 18.03.2013.* [Online] march 18, 2013. [Cited: may 15, 2013.] http://qeprize.org/.

2. **Internet World Stats.** Internet World Stats. *Internet Usage Statistics - The Internet Big Picture.* [Online] june 30, 2012. [Cited: may 14, 2013.] http://www.internetworldstats.com/stats.htm.

3. **Indvik, Lauren.** Mashable. *Global Internet Traffic Expected to Quadruple by 2015.* [Online] june 9, 2011. [Cited: may 10, 2013.] http://mashable.com/2011/06/09/global-internet-traffic-infographic/.

4. **Ortutay, Barbara.** USA Today. *Facebook to end Beacon tracking tool in settlement.* [Online] september 21, 2009. [Cited: may 19, 2013.] http://usatoday30.usatoday.com/tech/hotsites/2009-09-21-facebook-beacon_N.htm.

5. **Balogh, Zoltán.** *Do-Not-Track.* Balogh, Zoltán;, Budapest, Hungary : s.n., november 10, 2012.

6. **Barabási, Albert László.** *Villanások - a jövő kiszámítható.* Budapest : Helikon Kiadó Kft., 2010.

7. **Eckersley, Peter.** Electronic Frontier Foundation - Defending your rights in the digital world. *A Primer on Information Theory and Privacy.* [Online] january 26, 2013. [Cited: april 19, 2013.] https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy.

8. **D. Blondel, Vincent, et al.** Scientific Reports. *Unique in the Crowd: The privacy bounds of human mobility.* [Online] march 25, 2013. [Cited: may 4, 2013.] http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html.

9. **Zawadziński, Maciej.** Ad Technology and Analytics. *Alternatives to cookie tracking.* [Online] april 26, 2013. [Cited: may 12, 2013.] http://zawadzinski.com/2013/04/26/alternatives-to-cookie-tracking/.

10. **Boyle, Alan.** Cosmic Log. *Gay? Conservative? High IQ? Your Facebook 'likes' can reveal traits.* [Online] march 12, 2013. [Cited: may 03, 2013.] http://cosmiclog.nbcnews.com/_news/2013/03/11/17260270-gay-conservative-high-iq-your-facebook-likes-can-reveal-traits?lite.