Corvinus University of

Budapest

Phd in Business Informatics

PHD THESES

Zoltán Balogh

Collecting personal data and profiling on the Internet

Supervisor:
Dr. Péter Racskó
associate professor

Budapest, 2017

Department of Computer Science

PHD THESES

Zoltán Balogh

Collecting personal data and profiling on the Internet

Supervisor:
Dr. Péter Racskó
associate professor

Table of Contents

I. Research background and the justification of the topic

As the Internet has become part of our everyday lives, the development of web-based applications – mainly the browsers - have also accelerated. This has enabled websites to access more and more data about their visitors. By consuming any content online, the visitors' preferences and their browsing habits have become available to websites. Analyzing this data, the visitor can get personalized content, but at the same time their personal characteristics become revealed. As a result of personalized content, the phenomenon of "filter bubbles" has came to life, in which the user has access to content that matches his perceived properties, but has no control over what he can consume and what he can not. Due to user profiling, the initial anonymous web is now no longer anonymous.

As a result of technological advancement, the visitor preferences of the websites have become known, which has resulted in online recommendation systems. For example, the recommendation system of the US online broadcasting company – Netflix – influences the content consumption approximately 80% of their visitors'. (Carlos & Neil, 2015) That's why it is not surprising that the company is very interested in increasing the efficiency of the algorithms they use. Between 2006 and 2009, a $1 million Netflix Award was offered to those, who could improve the effectiveness of the so called recommendation system. The efficiency of the algorithm developed by the winning team surpassed that by 10.06%. (Lohr, 2009) Internet service providers are making great efforts to profile their customers accurately in order to have a competitive advantage.

In the field of information acquisition and comsumption, currently there is a paradigm shift: while previously users were searching for the information

that is relevant to them by using search engines, nowadays, besides using a search engine, online services are able to determine whether a particular content might be interesting to the visitor, by matching the visitor's preferences and the relevance of the contents. By using webpages (search pages, news portals, and social sites) that collect relevant informations of the visitors, then applying profiling algorithms on them, the visitors implicitly contribute to the collection of their detectable features. Based on this data, the website usually offers relevant content for their visitors' personality. In most cases, visitors do not even know that their personal data is being collected and there is no way to ignore this collection.

I.1. Researches

In my researches, I was searching for the set of available data of the visitors', that can be used to profile them. Also searched for the way to profile them. I have built up my researches as follows:

The goal of my first research is to examine what visitor data is available for the websites from a single domain. In this exploratory research, the personal characteritics of the visitors were extracted from the data of the software and hardware environment available to the Internet browsers. In the analysis phase, only trivial relationships between the visitors and their personal parameters were detected, personal characteristics were not found for visitors who did not disclose any information about themselves. I also investigated the **uncertainty-reducing power of parameters available to browsers**. The uncertainty-reducing power of the data collected of the visitors shows the probability to find the visitor within the sample.

The second research goal is to put the visitors of the Corvinus University of Budapest into **clusters based** on **their personal preferences** collected from the **social networks**, then to compare the clusters obtained with the clusters of data collected by the myPersonality Project (Stillwell & Kosinki, 2012).

The research shows how to extract personality traits from the data that is voluntarily provided by the visitors, that can be found in the "Like" database of the visitors. The personal data was anonymized before performing the research.

The active members of social networks can also express their interest in online content by clicking on the "Like" button. This visitor information is available on social networks. I analyzed the “Facebook Like” of the citizens of Corvinus University of Budapest with the help of a psychological API (Kielczewski, 2017), and then I clustered them with unsupervised learning methods based on their personal characteristics. The research also explores the differences between citizens of the Corvinus University of Budapest and the participants of myPersonality Project's personal clusters of personal qualities returned by the psychological API.

My third research goal was to find out **how to derive personal characteristics from non-personal characteristics of visitors**. By using the Apriori algorithm (Gautam, Ghodasara, & Parsania, 2014), I analyzed the behavior of the citizens of Corvinus University of Budapest in an e-learning environment to find out what personal traits can be deducted from their online behavior.

By clustering the visitors into groups based on their personality traits can be good for business, because targeted ads can be sent to each individual of the group. From the very beginning, one of Facebook's key business strategic

pillars is exploiting the potential of the social networking ad space. (Jeffrey, 2012) Visitors to websites can be categorized based on their business-critical features.

II. Used methodologies

I have developed my own data collection application for collecting the sample needed for the research. Before the development of the application, I thoroughly analyzed the applications that can be found online, as they did not meet all of the following criteria

- be able to save the widest range of available data
- provide low level of access to the collected data
- the known blocking applications should not be able to prevent their operation
- the application should not degrade the user experience in any way

The client-side part of the data collection application was written in Javascript, while the server-side part was developed in PHP, and the data was saved into MySQL database. After data collection part, the data was pre-processed with Pentaho. (Conversion of the GeoLocation to city and street name using Google Maps API, browser type extraction from HTTPUserAgent, assignment of the page visits to courses etc.)

Then I analyzed the personality traits of the citizens of Corvinus University of Budapest, obtained from their Facebook Likes (explicitly stated preferences via content sharing and likes). In the framework of the research it is possible to find out how personal data can be obtained from the data voluntarily provided by the visitors, their willingness to share the personal data, and the amount of personal data available online. The individuals were arranged into clusters based on their personality traits then compared with the clusters of the myPersonality Project research.

Researchers at the University of Cambridge Psychometrics Center set a milestone in online profiling in 2012, with 58,000 volunteer Facebook users who completed psychological tests and then their personal information on Facebook was analysed. After collating the results of the analyses, they developed an API available for everyone that can determine the psychological characteristics of the subject of the study, based on his likes on Facebook. (Kosinksi, Stillwell, & Graepel, 2013)

From the Facebook Likes, I acquired personality traits with the Apply Magic Sauce Psychological API and saved them in a database linked with the visitors' ID. Then I used the SPSS statistical package for my analyses, in addition to the basic statistical calculations, including K-means and hierarchical clustering methods.

In the research of „Psychological Characteristics of University Citizens from a Data Protection Point of view", I have used the personality traits extracted from Facebook Likes, the visitors' behavior in the online behavior and the properties of the visitors' software and hardware environment to conclude personality traits. In the sample, the non-personal variables are the elements of the attribute set, and the personal variables are the potential class variables. During the analysis, I used several classification algorithms, including the popular Apriori algorithm, which is able to search for association rules in large data sets. (Agrawal & Srikant, 1994) The Apriori algorithm was implemented in several data mining applications, and I used the Java-based Weka. (Witten, Frank, & Hall, 2011)

Apriori is an algorithm for frequent item set mining and association rule learning. It identifyes the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently

often in the database. The algorithm can effectively reduce the number of extracted rules, of which only those that are not less than a minimum support will be interesting. (Agrawal & Srikant, 1994) There are many different association rules for small samples, and the algorithm holds only those that apply to a relatively large number of copies.

III. Research results

During my research, I analysed websites accessible from a single domain and social networking sites or advertising agencies with links to other websites. The conclusions of the research and the reviewed literature for the following actors of the web browsing:

- **Individuals visiting websites**: relatively few information is only available for the single domain websites, from this amount of data, the personality trait extraction is limited. For social networking sites and advertising agencies, it is much more easier to extract personality traits since they have more data on their visitors'. The more accurate the visitor's personality traits are known, the more accurate content and advertisements can be sent to them, that might lead to the filter bubbles phenomenon. To avoid this, it is advisable to consciously use social networking sites, search engines and any other web sites, where advertising agencies have embedded advertising scripts.
- **Data Protection Authority:** the European Union Agency for Cyber Security, the European Union Agency for Network and Information Security (ENISA) - which was established in 2004 - makes recommendations and play a key role in shaping and implementing policy considerations for the European Union. (ENISA, 2017)
- **Software Developer / CIO:** To prevent the identification of the personality traits of the users/visitors, the applications should be prepared for the enhanced protections of the information they store. It is important to apply the "privacy by design" principle issued by ENISA and the frameworks based on it (ISO / IEC 29100) to prepare

applications for the protection of information during their design phase. (Danezis et al., 2015)

III.1. Characteristics of data, that can be collected from device browsers

The goal of this exploratory research is to show that the web browsers and tools available for visitors of websites are accessible from a single domain can be used to identify and track the visitors. I analyzed, to what extent the data available from the visitor's device browser can contribute to this.

During a single page load, a few hundred kilobytes of data is available for any website, from which valuable data can be extracted with data-mining methods, even personal information. This helps to identify and track individuals without using their personal information. The available parameters can be arranged into the following groups:

- features of the browser environment
- features of the software and hardware environment
- the hardware environment and browser features and
- constant parameters during an average session.

As a result, it turned out, that the more variables were used to identify the visitors, the greater the total uncertainty reducing power of all the variables used. This means that we can identify a device browser used by the visitor with a high certainty from the variables of a known population, provided that their parameters are known during the session. This also means that if we want to make it harder for our followers, it is advisable to use a proxy server for connecting the server, share as few parameters as possible with websites and third parties.

III.2. The Psychological characteristics of University citizens from privacy aspects

From the international data collected by the myPersonality Project and my own collection of data of the citizens of the Corvinus University of Budapest, I created clusters with unsupervised learning algorithms and compared the results. I found that the clusters are made up of individuals of similar personality traits. The data collected by the myPersonality Project were mainly citizens of the english speaking countries, and in my own research the participants were citizens of the Corvinus University of Budapest.

At the time of my research, the psychological analysis of Facebook likes can capture the following traits of individuals: Big5 personality traits, life satisfaction, intelligence, age, gender, sexual orientation, interest, political attitude, creed and family status. In the description of the Psychological API, the certainty (Pearson correlation coefficient) of the previously listed variables can be read. According to this description, the age, sex, sexual orientation, interests, political views, creed and marital status can be accurately predicted, while the Big5 life satisfaction and intelligence can only be predicted with a mediocre accuracy. I have run K-means and hierarchical clustering algorithms on both sets, resulting in exactly the same clusters, that's why they can be considered stable.

In the examined samples (myPersonality Project and Corvinus University of Budapest), two very similar clusters can be found:

- **Content publishers** with high openness, neuroticism and extroversion, who are unlikely to be in relationship with a high capability compromises
- **Conscientious individuals** are in serious relationship.

- The MyPersonality Project model shows a group of individuals who spend little time on the social network, presumably they are also present in the sample of Corvinus University of Budapest, but not detectable.

III.3. Retrieval of personal information from online data

The sample collected from the citizens of the Corvinus University of Budapest shows a correlation between the personality of the browser and the features of the software and hardware environment it uses and the visitor's online behavior.

I have searched for sub-sets of common element sets with the popular Aprior algorithm, then I examined those with high confidence and support. (Agrawal & Srikant, 1994) The algorithm has found rules with which the intelligence and life satisfaction of 10% of visitors can be predicted with high confidence.

These rules are only valid for the current set, and capable to demonstrate the potentials in the web-mining algorithms.

IV. Main references

Abramson, M., & Aha, D. W. (2013). User authentication from Web browsing behavior. *Florida Artificial Intelligence Research Society Conference* (old.: 6). St. Pete Beach: AAAI Press.

Agrawal, R., & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. *VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases* (old.: 487-499). San Francisco: Morgan Kaufmann Publishers Inc.

Andreas, P., & Marit, H. (2010. augusztus 10). *Privacy and Data Security, TU Dresden, Faculty of Computer Science.* Forrás: A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management: http://dud.inf.tu-dresden.de/Anon_Terminology.shtml

Barabási, A. (2010). *Villanások - a jövő kiszámítható.* Budapest: Helikon Kiadó Kft.

Bodon, F. (2010. február 28). Adatbányászati algoritmusok. Budapest, Magyarország.

Chris, H. J., Ashkan, S., Nathaniel, G., & Dietrich, W. J. (2012. january 1). Behavioral Advertising: The Offer You Can't Refuse. *Harvard Law & Policy Review vol. 6*, old.: 273-296.

Clarke, R. (1999). Internet Privacy Concerns Confirm the Case for Intervention. *Communications of ACM*, 60-67.

Cser, L., & Fajszi, B. (2004). *Üzleti tudás az adatok mélyén - Adatbányászat alkalmazói szemmel.* Budapest: Budapesti Műszaki és Gazdálkodástudományi Egyetem.

Danezis, G., Domingo-Ferrer, J., Hansen, M., Hoepman, J.-H., Le Métayer, D., Tirtea, R., & Schiffner, S. (2015. január 12). *European Union Agency for Network and Information Security.* Letöltés dátuma: 2017. május 14, forrás: Privacy and Data Protection by Design: https://www.enisa.europa.eu/publications/privacy-and-data-protection-by-design

Davenport, D. (2002. április). Anonymity on the Internet: Why the Price May Be Too High. *Communications of the ACM vol. 45, no. 4*, old.: 33-35.

Domokos, M. N. (2013). Az EU új adatvédelmi szabályozása – avagy „keep bangin' on the wall of Fortress Europe". *Jogi Fórum*, 1-46.

Eckersley, P. (2013, january 26). *Electronic Frontier Foundation - Defending your rights in the digital world*. Retrieved april 19, 2013, from A Primer on Information Theory and Privacy: https://www.eff.org/deeplinks/2010/01/primer-information-theory-and-privacy

ENISA. (2017). *European Union Agency for Network and Information Security.* Letöltés dátuma: 2017. május 14, forrás: About ENISA: https://www.enisa.europa.eu/about-enisa

Escobido, M., & Gillian, S. (2013). Can Personality Type be Predicted by Social Media Network Structures? *The Asian Conference on Psychology & the Behavioral Sciences.* Osaka: The International Academic Forum.

Európai Bizottság. (2015. július 11). *A személyes adatok védelme.* Forrás: Európai Bizottság honlapja: http://ec.europa.eu/justice/data-protection/index_hu.htm

France, B., & Robert, C. E. (2011). Privacy in the digital age: A review of information privacy research in information systems. *MISQ, volume 35, issue 4*, 1017-1041.

Haig, Z., Kovács, L., Ványa, L., & Vass, S. (2014). *Elektronikai hadviselés.* Budapest: Nemzeti Közszolgálati Egyetem.

Hunyadi, L., & Vita, L. (2006). *Statisztika közgazdászoknak.* Budapest: Központi Statisztikai Hivatal.

Jia-Ching, Y., Chu-Yu, C., & Vincent, T. S. (2012). Mining web navigation patterns with dynamic thresholds for navigation prediction. *IEEE Computer Society 2012* (old.: 614-619). Hangzhou: IEEE.

John, L., Manuel, B., & Luis, A. v. (2004. február). Telling Humans and Computers Apart Automatically. *Communications of the ACM*, 57-60. Letöltés dátuma: 2013. július 14, forrás: http://www.cs.cmu.edu/~biglou/captcha_cacm.pdf

Kang, R., Brown, S., & Kiesler, S. (2013). Why do people seek anonymity on the internet?: informing policy and design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2657-26666.

Kennedy, H. (2006). Beyond anonymity, or future directions for internet identity research. *New Media & Society, Vol 8, Issue 6*, 859-876.

Kiss, A. (2015. február 23). *Az adatokhoz, adatbázisokhoz kapcsolódó jogi szabályozás 1.* (A. Kiss, Előadó) Budapest.

Kosinksi, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *PNAS*, 5802-5805.

Kosinksi, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. In U. o.

Kenneth Wachter (Szerk.), *Proceedings of the National Academy of Sciences of the United States of America*. *110*, old.: 5802–5805. Berkeley: PNAS.

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., & Graepel, T. (2013. october 19). Manifestations of user personality in website choice and behaviour on online social networks. *Machine Learning, June 2014, Volume 95*, old.: 357-380.

Kosinski, M., Las Casas, D., Paulo Pesce, J., Quercia, D., Stillwell, D., Almeida, V., & Crowcroft, J. (2012). Facebook and Privacy: The Balancing Act of Personality, Gender, and Relationship Currency. *Sixth International AAAI Conference on Weblogs and Social Media.* Dublin: ICWSM.

Kovács, E. (2014). *Többváltozós adatelemzés.* Budapest: Typotex.

Nan, Z., Aaron, P., & Haining, W. (dátum nélk.). *An Efficient User Verification System via Mouse.*

Nemeslaki, A., Kis, G., Duma, L., & Szántai, T. (2004). *e-Business: Üzleti modellek.* Budapest: ADECOM Kommunikációs Szolgáltató Rt.

Peter, O., David, G., David, L., Warren, F., & Jonathan, N. B. (2005). Continuous Identity Verification. *Jur*, 20-24.

Racskó, P. (2012). A számítási felhő az Európai Unió Egén. *Vezetéstudomány*, old.: 1-16.

Shababi, C., Zarkesh, M. A., Adibi, J., & Shah, V. (1997). Knowledge discovery from users web page navigation. *26th IEEE International Conference on research in Data Engineering*, (old.: 20-29).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 379-423.

Stillwell, D. J., & Kosinki, M. (2012). *myPersonality project: Example of successful utilization of online social networks for large-scale social research.* Cambridge, University of Cambridge, UK: The Psychometrics Centre.

Stillwell, D., Kosinki, M., Rust, J., & Wang, N. (2012. february 3). Can Well-Being be Measured Using Facebook Status Updates? Validation of Facebook's Gross National Happiness Index. *Social Indicators Research vol 115, issue 1*, old.: 483-491.

Szabó, A. (2010). *Random Forests - Véletlen erdők.* Letöltés dátuma: 2017. január 8, forrás: Adatbányászat és Keresés Csoport: https://dms.sztaki.hu/sites/dms.sztaki.hu/files/file/2011/randomforests.pdf

Személyes adatok feldolgozása vonatkozásában az egyének védelméről és az ilyen adatok szabad áramlásáról, 95/46/EK (Az Európai Parlament és a Tanács 1995. október 24).

Voulodimos , A. S., & Patrikakis , C. Z. (2009. december). Quantifying privacy in terms of entropy for context aware services. *Identity in the Information Society, 2*(2), 155-169.

Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining - 3rd edition.* Burlington: Morgan Kaufmann.

Youyou, W., Kosinki, M., & Stillwell, D. (2015. január 27). Computer-based personality judgments are more accurate than those made by humans. *PNAS*, old.: 1036-1040.

V. List of publications related to the topic

Scientific books, book parts

Balogh Zoltán [2012]: Internetes anonimitás aktorai. In *Gazdaság, Társadalom II. A jövő és 2012*, *Arisztotelész*, ISBN: 987-963-87721-8-3, pp. 6-16.

Referenced papers in Hungarian

Racskó Péter, Szommer Károly, Balogh Zoltán [2014]: *Az online világban hagyott virtuális lábnyomokban rejlő információ és azok veszélyei*. In *Vezetéstudomány, volume XLV., issue 2014. 7-8.*, *BCE*, ISSN: 0133-0179, pp. 97-104.

Szommer Károly, Balogh Zoltán [2016]: *Geotagging használata Magyarországon*. In Minőség és megbízhatóság, *EOQ MNB Egyesület*, ISSN: 0580-4485, pp. 140-147

Referenced papers in English

Balogh Zoltán [2012]: Identification in eLearning Environment. In SEFBIS Journal 2013 No.8, ISSN: 1788-2265, pp. 81-86

Rétallér Orsolya, Balogh Zoltán [2015]: *Specialities of Psychological Traits of Citizens of Corvinus University of Budapest*. In *Hadmérnök X. Évfolyam 4. szám*, ISSN 1788-1919, pp. 193-204

Other (english)

Balogh Zoltán [2012]: *Anonymity over the internet*. Proceeding of *Cyter conference 2012*, ISBN: 978-80-01-05072-9, pp. 1

Balogh Zoltán [2012]: *Anonymity over the internet*. Proceeding of *18th International ICE-Conference on Engineering, Technology and Innovation*, ISBN: 978-1-4673-2275-1, pp. 377-383

Balogh Zoltán [2012]: *Do-not-track*. Proceeding of *Professzorok az Európai Magyarországért 2012*, ISBN: 978-963-88433-7-1, pp. 9

Balogh Zoltán [2012]: *Potential dangers of using the web*. Proceeding of *GIKOF e-Journal*, pp. 8

Szommer Károly, Balogh Zoltán [2015]: *Dangers of sharing platforms used by people of different personalities*. Proceeding of *ACOMP 2015, IEEE Computer Society*, ISBN-13: 978-1-4673-8234-2, pp. 7-11.

Balogh Zoltán [2016]: *Mining web data with Apriori algorithm*. Proceeding of *SKIMA 2016*, ISBN: 978-1-5090-3298-3, pp. 10