eLaw
Working Paper Series

No 2018/003 - ELAW– 24 April 2019

Data Mining and Profiling in Big Data

Bart Custers

Universiteit Leiden
eLaw

The SAGE Encyclopedia of Surveillance, Security, and Privacy

Data Mining and Profiling in Big Data

Contributors: Bart Custers
Edited by: Bruce A. Arrigo
Book Title: The SAGE Encyclopedia of Surveillance, Security, and Privacy
Chapter Title: "Data Mining and Profiling in Big Data"
Pub. Date: 2018
Access Date: April 25, 2018
Publishing Company: SAGE Publications, Inc.
City: Thousand Oaks
Print ISBN: 9781483359946
Online ISBN: 9781483359922
DOI: http://dx.doi.org/10.4135/9781483359922.n121
Print pages: 277-279

Data mining and profiling are technologies used for analyzing and interpreting large amounts of data (a set of facts) to obtain knowledge (patterns in the data that are interesting and certain enough for a user). In the information society, vast amounts of data are collected, stored, and processed by both public and private organizations. When dealing with large data sets, particularly in the context of big data, human intuition may be insufficient to obtain insight into or an overview of the data available.

Data mining and group profiling are considered separate technologies, even though they are often used together. Whereas the focus of data mining is on finding novel patterns and relations in data sets, the focus of profiling is on ascribing characteristics to individuals or groups of people. Profiling may be carried out without the use of data mining, and vice versa. In some cases, profiling may not involve (much) technology—for instance, when psychologically profiling a serial killer.

Advantages and Disadvantages

Profiles may offer general advantages, such as enabling the selection of target groups, customization, and cost efficiency. For corporations, profiles may be useful to identify new customers, personalize special offers, evaluate the profitability of product groups, and assess credit scores. Particularly, banks and insurance companies are interested in risk profiles to determine to whom to provide loans, mortgages, and insurances and under which conditions. For government agencies, profiles may be useful to identify target groups for their policies, to evaluate their policies, and to optimize public services. Particularly, criminal investigation organizations, including police agencies, and intelligence organizations are interested in risk profiles to identify criminals and terrorists, to assess and predict where crime will take place (so-called hotspots), and to disclose criminal networks.

General disadvantages of group profiles may involve, for instance, unjustified discrimination (e.g., when profiles contain sensitive characteristics like ethnicity or gender, which are used for decision making), stigmatization (when profiles become public knowledge), dehumanization (regarding people as data sets rather than human beings), de-individualization (regarding people as parts of groups rather than unique individuals), loss of privacy (when predicting characteristics that people do not want to disclose), loss of autonomy (as data mining and profiling practices may not be very transparent), and being confronted with unwanted information (e.g., with life expectancies). Many of the effects of group profiles may be considered advantageous as well as disadvantageous, depending on the context and the way in which, and by whom, the group profile is used.

Data Mining

Data mining is an automated analysis of data, using mathematical algorithms, to find new patterns and relations in (large amounts of) data. Data mining is a step in a process called knowledge discovery in databases. Knowledge discovery in databases is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. This process consists of five successive steps: (1) data collection, (2) data preparation, (3) data mining, (4) interpretation, and (5) determining actions. Hence, the third step is the actual data mining stage, in which the data are analyzed to find certain patterns or relations. This is done using mathematical algorithms. Data mining is different from traditional database techniques or statistical methods because what is being looked for does not

necessarily have to be known. Thus, data mining may be used to discover new patterns or to confirm suspected relationships. The former is called a bottom-up or data-driven approach, because it starts with the data and then theories based on the discovered patterns are built. The latter is called a top-down or theory-driven approach, because it starts with a hypothesis and then the data are checked to determine whether they are consistent with the hypothesis.

There are many different data mining techniques. The most common types of discovery algorithms with regard to group profiling are clustering, classification, and, to some extent, regression. Clustering is used to describe data by forming groups with similar properties; classification is used to map data into several predefined classes; and regression is used to describe data with a mathematical function.

In data mining, a pattern is a statement that describes relationships in a (sub)set of data such that the statement is simpler than the enumeration of all the facts in the (sub)set of data. When a pattern in data is interesting and certain enough for a user, according to the user's criteria, it is called knowledge. Patterns are interesting when they are novel (which depends on the user's knowledge), useful (which depends on the user's goal), and nontrivial to compute (which depends on the user's means of discovering patterns, e.g., the available data and the available people and/or technologies to process the data). For a pattern to be considered knowledge, a particular certainty is also required. A pattern is not likely to be true across all the data. This makes it necessary to express the certainty of the pattern. Certainty may involve several factors, such as the integrity of the data and the size of the sample.

The knowledge discovered may concern people, in which case it may result in profiles. These profiles may concern individuals, resulting in individual profiles, or they may concern groups, resulting in group profiles. When the knowledge reveals the probabilities of particular characteristics of individuals or groups, the profiles are generally referred to as risk profiles.

Profiling

Profiling is the process of creating profiles—that is, a property or a collection of properties of an individual or a group of people. Although profiles can be made of many things, such as countries, companies, or processes, in the context of surveillance, security, and privacy the profiles of people or groups of people are most relevant. Personal profiles are also referred to as individual profiles or customer profiles, while group profiles are also referred to as aggregated profiles.

A personal profile is a property or a collection of properties of a particular individual. A property or characteristic is the same as an *attribute*, a term more often used in the computer sciences. An example of a personal profile is the profile of Mr. John Smith (age 47 years), who is married, has three children, earns $75,000 a year, has two credit cards, and has no criminal record. He was fined for speeding twice last year and was hospitalized once in his lifetime, last year, because of appendicitis.

A group profile is a property or a collection of properties of a particular group of people. Group profiles may contain information that is already known—for instance, people who smoke live, on average, fewer years than people who do not. But group profiles may also reveal new facts —for instance, people living in zip code area 90003 may have a (significantly) larger than average chance of having asthma. Group profiles do not have to describe a causal relation. For instance, people driving red cars may have (significantly) more chances of getting lung cancer than people driving blue cars. Note that group profiles differ from individual profiles

with regard to the fact that the properties in the profile may be valid for the group and for individuals as members of that group but not for those individuals as such. This is referred to as non-distributivity or a nondistributive profile. When the properties in a profile are valid for each individual member of a group as an individual, it is referred to as distributivity or a distributive profile.

Several data mining methods are particularly suitable for profiling. For instance, classification and clustering may be used to identify groups. Regression techniques may be useful for making predictions about a known individual or group.

See also Big Data; Passenger Profiling; Privacy; Social Network Analysis; Technology

- data mining
- profiling
- big data
- mining
- algorithms
- property
- regression

Bart Custers
http://dx.doi.org/10.4135/9781483359922.n121
10.4135/9781483359922.n121

Further Readings

Bygrave, L. A. Data Protection Law.New York, NY: Kluwer Law International, 2002.
Custers, B. H. M. The Power of Knowledge: Ethical, Legal, and Technological Aspects of Data Mining and Group Profiling in Epidemiology.Tilburg, Netherlands: Wolf Legal, 2004.
Custers, B. H. M. et al., eds. Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases.Heidelberg, Germany: Springer, 2013.
Fayyad, U. M., et al. “The KDD Process for Extracting Useful Knowledge From Volumes of Data.” Communications of the ACM, v.39/11 (1996).
Harcourt, B. E. Against Prediction: Profiling, Policing and Punishing in an Actuarial Age.Chicago, IL: University of Chicago Press, 2007.
Hildebrandt, M. and S. Gutwirth. Profiling the European Citizen.Heidelberg, Germany: Springer, 2008.
Mayer-Schönberger, V. and K. Cukier. Big Data: A Revolution That Will Transform How We Live, Work and Think.New York, NY: Houghton, Mifflin, Harcourt, 2013.
Schauer, F. Profiles, Probabilities and Stereotypes.Cambridge, MA: Harvard University Press, 2003.
Solove, D. The Digital Person: Technology and Privacy in the Information Age.New York: New York University Press, 2004.
Zarsky, T. Mine Your Own Business! Yale Journal of Law and Technology, v.5 (2003).