Data Mining and Profiling

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

Data mining and group profiling are often mentioned in the same breath, but they may be considered separate technologies, even though they are often used together. Whereas the focus of data mining is on finding novel patterns and relations in datasets, 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.

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, evaluating profitability of product groups, and assessing credit scores. Particularly banks and insurance companies are interested in risk profiles to determine whom to provide loans, mortgages and insurances to and under which conditions. For government agencies profiles may be useful to identify target groups for their policies, evaluate their policies and optimize public services. Particularly criminal investigation organisations, including police agencies, and intelligence organisations 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 (for instance, when profiles contain sensitive characteristics like ethnicity or gender for decision-making), stigmatization (when profiles become public knowledge), de-humanization (regarding people as datasets 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 confrontation with unwanted information (for instance, with life expectancies). It should be noted, however, that 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, in order to find new patterns and relations in (large amounts of) data. Data mining is a step in a process called Knowledge Discovery in Databases (KDD). 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: data collection, data preparation, data mining, interpretation and determining actions. Hence, the third step is the actual data-mining stage, in which the data are analyzed in order 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 is checked to determine whether it is 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, this 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, such as 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 discovered knowledge may concern people, in which case they 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 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. Although profiles can be made of many things, such as countries, companies or processes, in the context surveillance, security and privacy the profiles of people or groups of people are most relevant. A profile is a property or a collection of properties of an individual or a group of people. 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 computer sciences. An example of a personal profile is the profile of Mr. John Smith (47), who is married, has three children, earns 75,000 euro a year, has two credit cards and no criminal record. He was fined for speeding twice last year and was hospitalized once in his life, 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, a few years less than people who do not. But group profiles may also reveal new facts, for instance, people living in zip code area 5439 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 non-distributive profiles. When properties in a profile are valid for each individual member of a group as an individual, this is referred to as distributivity or distributive profiles.

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.

Bart Custers
Leiden University, The Netherlands

See Also: Big Data; Privacy; Passenger Profiling; Racial Profiling; Social Network Analysis; Technology.

Further Readings

Bygrave, L.A. (2002) *Data Protection Law*. New York: Kluwer Law International.

Custers, B.H.M. (2004) *The Power of Knowledge; Ethical, Legal, and Technological Aspects of Data Mining and Group Profiling in Epidemiology*. Tilburg: Wolf Legal Publishers.

Custers, B.H.M., Calders, T., Schermer, B., and Zarsky, T. (eds.) (2013) *Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases*, Heidelberg: Springer.

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM*, vol. 39, nr. 11.

Harcourt, B.E. (2007) *Against Prediction: Profiling, Policing and Punishing in an Actuarial Age*. Chicago: University of Chicago Press.

Hildebrandt, M., and Gutwirth, S. (2008) *Profiling the European Citizen*, Heidelberg: Springer.

Mayer-Schönberger, V. and Cukier, K. (2013) *Big Data: A Revolution That Will Transform How We Live, Work and Think*, New York: Houghton, Mifflin, Harcourt Publishing Company.

Schauer, F. (2003) *Profiles, Probabilities and Stereotypes*, Cambridge: Harvard University Press.

Zarsky, T. (2003) Mine Your Own Business! *Yale Journal of Law and Technology* 5, 57.

Solove, D. (2004) *The Digital Person: Technology and Privacy in the Information Age*, New York: New York University Press.

The research leading to the presented results has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 731873.