Methods and Techniques of Data Mining

6

Ana Funes
Universidad Nacional de San Luis, Argentina

Aristides Dasso
Universidad Nacional de San Luis, Argentina

INTRODUCTION

"Can machines think?", this is the question that Turing (1950) put forward already in 1950. It is possible to consider the discussion in that article as a starting point in history of what nowadays is called Artificial Intelligence (AI), although, previous to that time, there are some works that could be considered as what is now called AI.

Already in antiquity, philosophers and thinkers in general had advanced the idea that "mechanical men" could be constructed and they would be able to mimic human beings in action and thought. So it could be said that the concept of artificial intelligence has existed for a very long time previous to 1950, the approximate date considered here as the start of what is now referred as AI.

AI has had from the 50s a fruitful history. Along the years several branches have emerged such as Machine Learning, Data Mining, Robotics, Statistical Learning, Deep Learning and others. Research on the subject and the development of tools and methods have grown significantly, specially over the past twenty years. The constant search for the causes of diseases, the improvement of automatic diagnoses methods, financial data analytics and market tendencies, among others, are only some of the innumerable applications where analysis and discovery of new patterns have fuelled the research and development of new methods, all related to Machine Learning, knowledge extraction from what is now being called Big Data, Knowledge Discovery in Databases or KDD, and Data Mining. This chapter focuses on Data Mining methods and algorithms as approaches to different learning tasks. The relationship of Data Mining with other disciplines inserted in the area of Artificial Intelligence is also analyzed, as well as a brief discussion on issues associated to the discipline and future research directions are also presented.

BACKGROUND

Theoretical discussions about the possibility of creating automatons or supposedly intelligent machines have been coming for a long time. In Ernst & Newell (1969) a research was conducted on the development of a computer program with general problem-solving capabilities. The program could solve problems like the Tower of Hanoi or prove theorems in first-order predicate calculus.

Weizenbaum, J. (1976), author of a then well-known computer program initially called ELIZA and later DOCTOR that allowed parodying the role of a Rogerian psychotherapist, was surprised, he says, for three events: (i) A number of psychiatrists thought that the program could become an automatic form of psychotherapy. (ii) How quickly and deeply people became emotionally involved with the program. (iii) The dissemination of the belief that the program showed a general solution to the problem of understanding natural language by computers. This situation reported by Weizenbaum can be compared

DOI: 10.4018/978-1-7998-3479-3.ch051

with The Turk, the well-known chess player who for decades, from the end of the 18th century until the end of the 19th century, was believed to be a machine capable of playing chess. This, and many others around the same time, can be predecessors of programs that nowadays are considered to be part of the general field of AI.

In his textbook Wiston (1977) says that "The central goals of AI are to make computers more useful and to understand the principles which make intelligence possible".

Many authors agree that AI is an important knowledge area that is closely related to other disciplines. As Russell, S. J. & Norvig, P. (2010) express, "Artificial Intelligence (AI) is a big field ... which encompasses logic, probability, and continuous mathematics; perception, reasoning, learning, and action; and everything from microelectronic devices to robotic planetary explorers". AI comprises not only Machine Learning but other disciplines such as Robotics, Logic Programming and Inductive Logic Programming (ILP), a field of Machine Learning and Data Mining. These different areas could conceivably be either form part of AI or have sprout from it.

AI has had different goals along its history. Initially the idea embodied in Turing's question was paramount. Could a computer be built that acts like a human being?, hence, could it solve the problems that a human could? Several tests were developed to assess such ability. Turing (1950) put forward what is now called the Turing Test, but that he called the 'The Imitation Game'.On the Turing Test there are different opinions. Cohen, Paul R. (2006) summarizes them saying that there are three kinds of arguments against Turing's test, namely some say that the test is irrelevant, it should be forgotten, get rid of it, however it still stands today. There are competitions, variations on the test, etc. Another argument against the test is related to the philosophy connected to the question of whether machines can think, or what is thinking and if machines really can do it or is it just a simulacrum. Others give a methodological argument when asking the question. is the Imitation Game, −as Turing put forward− a valid methodological means of definitely determining whether a machine can think?, "They argue that the test is methodologically flawed and is based in bad philosophy, that it exposes cultural biases and naïveté about what Turing calls the "programming" required to pass the test." Cohen, Paul R. (2006, p. 61). For more on the Turing test see Hernández-Orallo, J. (2000).

AI has greatly evolved since Turing's paper times. That evolution has boosted the development of different branches, stimulated by practical concerns such as pattern recognition, data analysis and knowledge discovery. One of these areas is Data Mining, in which this chapter is focused.

The development of some AI fields has benefited from the existence of large volumes of data proceeding from the most diverse sources and domains. The Knowledge Discovery in Databases (KDD) process and methods of Data Mining allows for the discovery of knowledge in data that is hidden to humans, particularly when data volumes are large or even extremely large. These methods present the knowledge extracted under different ways: rules, equations, decision trees, etc., and help in answering questions such as what are the groups from a population of individuals with common characteristics?, is a client reliable?, is an e-mail spam?, etc.

Answers to these kind of questions are different from those obtained from traditional queries in On Line Transactions Processing (OLTP), where the information is not hidden neither is discovered, but it is summarized in an agreed format or report. They also differ from information proceeding from Online Analytical Processing (OLAP), which can be presented in different perspectives or aggregated in different ways and not just summarized as in OLTP, and that can even escalate to the use of big data, where OLTP fails. However, both of these methods are not capable of discovering new knowledge neither producing new patterns nor rules as the KDD process permits.