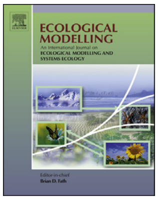
Ecological Modelling 295 (2015) 5–17

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel



Data mining and linked open data – New perspectives for data analysis in environmental research

Angela Lausch [*], Andreas Schmidt, Lutz Tischendorf

Department of Computational Landscape Ecology, Helmholtz Centre for Environmental Research – UFZ, Permoserstr, 15/D-04318, Leipzig, Germany

ARTICLE INFO

Article history:
Available online 30 September 2014

Keywords:
Data Mining
Knowledge Discovery Process
Data Mining Tools
Linked Open Data
Semantic
RapidMiner

ABSTRACT

The rapid development in information and computer technology has facilitated an extreme increase in the collection and storage of digital data. However, the associated rapid increase in digital data volumes does not automatically correlate with new insights and advances in our understanding of those data. The relatively new technique of data mining offers a promising way to extract knowledge and patterns from large, multidimensional and complex data sets. This paper therefore aims to provide a comprehensive overview of existing data mining techniques and related tools and to illustrate the potential of data mining for different research areas by means of example applications. Despite a number of conventional data mining techniques and methods, these classical approaches are restricted to isolated or "silo" data sets and therefore remain primarily stand alone and specialized in nature. Highly complex and mostly interdisciplinary questions in environmental research cannot be answered sufficiently using isolated or area-based data mining approaches. To this end, the linked open data (LOD) approach will be presented as a new possibility in support of complex and inter-disciplinary data mining analysis. The merit of LOD will be explained using examples from medicine and environmental research. The advantages of LOD data mining will be weighed against classical data mining techniques. LOD offers unique and new possibilities for interdisciplinary data analysis, modeling and projection for multidimensional, complex landscapes and may facilitate new insights and answers to complex environmental questions. Our paper aims to encourage those research scientists which do not have extensive programming and data mining knowledge to take advantage of existing data mining tools, to embrace classical data mining and LOD approaches in support of gaining more insight and recognizing patterns in highly complex data sets.

1. Introduction

The rapid technological advances in information technology of the 21st century are intrinsically linked to a gigantic increase in data and information. This is the result of increased networking and globalization, continuous improvement in computer engineering, media storage, highly sophisticated data bases, the Internet as a platform for communication and of the enormous expansion of automated data collection via sensors, monitoring systems and mobile and smartphone applications. The ever smaller-scale automated measuring, cataloging and monitoring of so many areas of people's lives as well as their social environment and networks leads to a "datafication" of the world (Kreissl, 2013). Rekow (2013) refers to the phenomenon of growing data volume in all areas of life also as "information overload". Schmid (2013) points out that with the increasing networking and ever-growing computing capacity we find ourselves in a "profound transformation process" of the 21st century where more and more "decision-making processes have to be outsourced to technical systems" in order to be at all able to function. Internet services like Facebook, Google and Wolfram Alpha aim "to make all systematic knowledge immediately computable and accessible to everyone" (Alpha, 2013).

Estimates suggest that the entire amount of data in the world doubles every 20 months (Runkler, 2010). According to Meyer and Lüling (2003), researchers at the University of Berkeley have calculated that around 1 exabyte (= 1 million terabytes) of data is generated every year. And so by the end of 2015, the annual Internet data traffic will amount globally to 1 zettabyte (1 zettabyte = 1000 exabytes). For the sake of comparison, "it would take over five years to watch the amount of videos which will be transferred per second through the Internet by the year 2015" (Computer, 2013). Whereas in the 1990s there was still a deficit in the availability of digital data, the relation between data

[*] Corresponding author. Tel.: +49 3412351961.
E-mail address: angela.lausch@ufz.de (A. Lausch).

http://dx.doi.org/10.1016/j.ecolmodel.2014.09.018
0304-3800/© 2014 Elsevier B.V. All rights reserved.

availability and existing evaluation methods and algorithms has changed completely.

Nowadays, generating data does not mean insight. Only a fraction of the data is used to gain insight. As always, the statement applies that "data is not insight". Those who have a lot of data "have possibilities to gain insight, but those who have the right analytical tools and instruments, also hold the key to acquiring insight". While large amounts of available data have, for the most part, been archived and stored, only small parts have really been analyzed, used and understood and processed in human-understandable ways (Begum, 2013).

In order to understand data and to gain insight, the data has to be sorted, transformed, harmonized and processed both statistically and analytically. The potential growth increase in data bases in all areas makes the analysis of huge amounts of more complex information more difficult and less clear. Using the procedures of classical statistics that have been common up to now, it is becoming apparent that very limited information and knowledge can be gleamed and insufficient progress made from the huge and complex amounts of data. Only if future research succeeds in extracting information from this complex amount of data can this be the basis for good operative and strategic decisions and projections.

In recent years, an enormous development in the area of complex data analysis has been made with data mining. Data mining, often termed knowledge discovery in databases KDD, is defined as the "non-trivial process of identifying valid, new and potentially useful and understandable patterns in data bases" (Saake and Heuer, 1999). Data mining is a rather new field in computer science that pursues the objective of extracting knowledge and analyzing complex data to find existing associations, to extract structures, patterns and regularities in large and complex data bases (Witten et al., 2011; Joseph et al., 2013). Currently, there are a number of common data mining techniques to analyze texts, web contents, images, pictures, videos and spatial data. Compared to classical statistical techniques, all of these data mining techniques exhibit very good results when it comes to discovering and analyzing patterns and coherencies of the researched data.

Environmental research is an interdisciplinary research field that is complex, multifaceted and very dynamic. In this field, data mining holds a particular promise to gain true insight by means of unveiling associations and relationships. But at the moment, data mining methods are only used in some areas. And here, too, an interdisciplinary analysis approach is not given sufficient support. LOD is a new development in support of internet based data mining where complex, highly dynamic and, for the first time, interdisciplinary data mining analyses are becoming possible. The goal of this paper is: (1) to provide an overview of existing data mining techniques, tools and methods, (2) to show the potential data mining holds for research using examples, (3) to give an overview of the limitations and necessary requirements of data mining for interdisciplinary environmental research, (4) to present the LOD approach as a new possibility in complex and interdisciplinary data mining research, (5) to encourage non-mathematicians and non-computer scientists from all scientific disciplines who do not have programming knowledge to use these novel tools in their research.

2. Data mining approaches

Data mining deals with the analysis, recognition and establishment of associations and patterns in existing data. And so data mining defines itself as a process to identify patterns in data with the potential to make non-trivial projections about as yet unknown patterns in data (Witten et al., 2011).

2.1. Data mining vs. statistical approaches

Data mining and conventional statistical analyses have different purposes. Whereas classical statistical approaches focus primarily on verifying stated hypotheses, data mining methods search through many possible, mostly unknown hypotheses (Witten and Eibe, 2001). Coupling statistical and data mining methods will be the only way to gain insight and knowledge from the ever increasing amount of digital data.

As Witten et al. (2011) pointed out, analyzing diverse and complex data in the future will not only require a coupling of data mining and statistical methods but the merging of disciplines and methods such as pattern recognition, data bases, artificial intelligence and machine learning algorithms (Fig. 1).

2.2. Data mining process

Data mining involves the entire process from the provision of data right up to the projection and application of model findings to new, unknown data structures. This process includes (a) techniques to preprocess data, (b) the actual data mining system (DM system) and c) interpreting and evaluating data.

Necessary preprocessing steps for data mining include data selection, preprocessing and transforming data into suitable data formats. The DM system itself is at the core of the actual data mining process, which is made up of three Phases: Training, Test and Validation (Fig. 2). Within this process the objective of data mining is to repeatedly attempt to determine an estimated value (e.g., the buying behavior of a customer) based on the researched data (e.g., market data), which is compared with a predetermined reference value (target variable) (Witten and Eibe, 2001). This process is repeated iteratively until the comparison of estimated and reference values results in an acceptable value. This obtained model now forms the basis for Phase 3 of the interpretation and evaluation as well as for deriving knowledge based on other, as yet unknown data.

2.3. Types of data mining

Data mining systems can be categorized depending on their objectives. Fig. 2 provides an overview.

Data Mining: Data mining includes the analysis of numeric and categorical data in large and complex data sets. Often this term is used to generally describe more specialized techniques, such as text, web or spatial data mining.

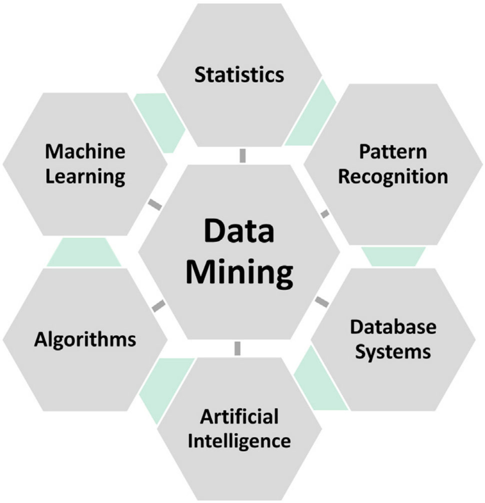


Fig. 1. Confluence of different multiple disciplines in the data mining process.