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.

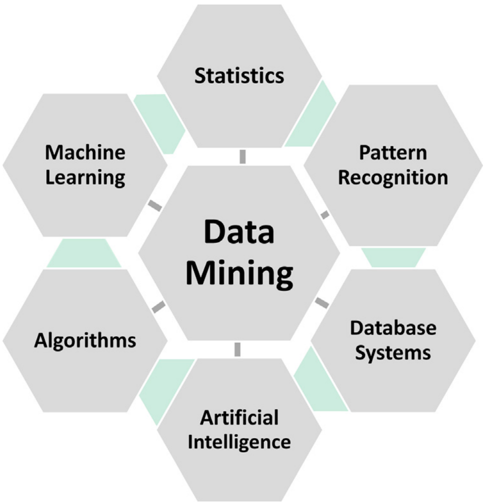

Fig. 1. Confluence of different multiple disciplines in the data mining process.

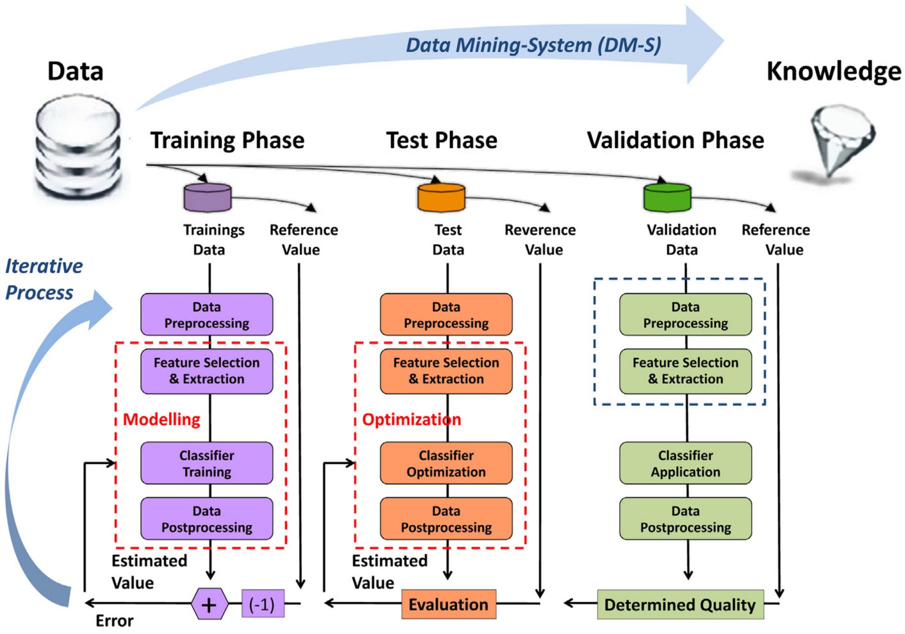

Fig. 2. Data mining process with the data mining system (DM-S) in the Phases (1) Training Phase, (2) Test Phase and (3) Validation Phase. The data mining process works in a comparable way in all types of data mining types like text mining or web mining (changed according to Fayyad et al., 1996 and Tanner, 2013).

Text Mining: Text mining includes algorithms to analyze lexical and grammatical aspects of texts (Dörre et al., 2001). In text mining algorithms the actual text is broken down into specific structures whereby patterns and core information of the text are recorded and grouped and classified using data mining methods. The first text mining tools could record contents and structures of simple text documents like Microsoft Word and Acrobat PDF documents. New developments in text mining now scan and analyze unstructured text in s, memos, surveys, chats, notes, white papers, forums and presentations (Begum, 2013).

Web-Mining: According to Bensberg and Weiß (1999) web mining is understood to be the application of data mining methods on information compiled on the Internet. In web mining, a distinction is made between (1) web content mining, which is the analysis of website contents; (2) web structure or relationship mining, i.e. the analysis of in- and outbound hyperlinks of websites and (3) web usage mining, which records and analyses user interaction with websites by scanning log files.

Image-Mining: The goal of image mining techniques is to analyze and extract spatial patterns in image data which are not explicitly stored in the images (Mohanty et al., 2013). Pattern extraction is done in a variety of ways based on recognizing the existence and distribution of colors, texture, shape, distances and intensities in the image data.

Mining of Picture, Video data and Music: Pattern mining techniques are used more and more to recognize features in pictures, videos and music data. This involves, on the one hand, detecting disturbances in pictures, video and music files, but also using similarities in data. Furthermore, data mining is the only way to handle the huge amounts of complex data which is generated, for instance, on Google Image or YouTube. Of prime importance is the quick activation of content-based images/videos retrieval, indexing, classification and tracking.

Mining of Time Series data: In the analysis of time series, temporal relations in time data are researched by means of a special distance function like dynamic time warping. The goal is to recognize similarities over the course of the time series, even when these similar features are time-shifted over the course. A good example is "Google Correlate" (http://www.google.com/trends/correlate).

Spatial Data Mining (SDM): Spatial data mining pursues the goal of discovering patterns in large, multi-dimensional spatial data sets as created by remote sensing techniques in earth observation. Shekhar et al. (2010) point out that extracting interesting patterns and associations from complex and multi-dimensional spatial data bearing spatial dependencies and spatial-temporal autocorrelations is more difficult than from traditional numeric and categorical data.

If, in addition to spatial data, additional time series are incorporated into the pattern analysis, the term spatial-temporal data mining is used to describe this.

2.4. Methods of data mining

There are various methods in support of data mining, which can be grouped according to the objective in question. Here we provide an overview over the most commonly used data mining methods. We encourage the reader to also refer to Witten et al. (2011) and Chu (2014) for a comprehensive overview of the various methods available for data mining.

Association and sequence analysis: With the help of association analysis, relationships between objects can be made and quantified. Using specific indicators, which, in most cases, include the support, confidence and lift values, the strength of the identified associations can be evaluated.

Grouping and clustering: These are procedures which pool similar objects into groups with the goal of having very similar objects within one group and having the groups differ as much as possible from each another. Numerous metrics are used for characterizing similarities in groups.

Regression: With the assistance of regression analysis, functional dependencies are determined among the variables within a data set. Regression models are used to estimate or predict variables. To represent dependencies there are also linear and non-linear (e.g., square, logistic or Poisson) regression approaches in addition to linear ones (Witten et al., 2011).

Classification: The goal of classification is to find functions and models with whose help data objects can be assigned to previously identified classes. Deriving a model is based usually on a set of objects for which the respective class allocation is known. With the help of a model that has to be determined, objects are to be classified that have no known classification. The models can be determined with the help of neuronal networks, discriminant analyses or decision trees (Breiman et al., 1984) and random forest (Chu, 2014). Naïve Bayes methods as well as support vector machines (SVM) can also be used to classify data (Witten et al., 2011; Chu, 2014).

There are a number of other procedures such as time series analysis or visualization und evolutionary algorithms. Witten et al. (2011) as well as Goele and Chanana (2012) provide a comprehensive overview of these.

2.5. Tools for data mining

In recent years, data mining tools have become much more available, interactive and explorative. This development has been pioneered by systems, which enable scientists from different disciplines without specialized expertise in mathematics and computer programming to conduct data mining in their research areas.

The report of the 14th annual KDnuggets Software Poll (KDnuggets Annual Software Poll, 2013) reveals the use of open source and commercial data mining tools during 2012 and 2013 (Fig. 3). According to this poll, the open source data mining tools RapidMiner and R have been the most frequently used tools. The report also depicts that 29% of voters only used commercial software, 30% used open source software and 41% used both commercial and free software.

3. Open source tools for data mining

There are some capable open source data mining tools available with good application potential in environmental and interdisciplinary research even for users without extensive IT and computer programming skills. Hirudkar and Sherekar (2013) provide a comparative analysis of these data mining tools and techniques. The four most commonly used open source data mining tools are RapidMiner, R, WEKA and KNIME. They are described briefly below:

RapidMiner: RapidMiner (http://www.rapidminer.com) is an open source framework which supports various types of data mining such as text, web, image or linked open data mining. Through its sophisticated graphical user interface, data mining processes can be implemented and executed quickly, intuitively and without computer programming knowledge (Fig. 9). Furthermore, RapidMiner offers many necessary pre- and post-process stages in support of data integration, extract, transform and loading (ETL), analysis and reporting in one single and coherent graphical user interface (Hofmann and Klinkenberg, 2013). RapidMiner furthermore offers a comprehensive selection of data mining operations as well as an integration of the WEKA and R libraries. RapidMiner is implemented in Java and can integrate processes developed by users as plug-ins. Moreover, it also offers an application programming interface (API), through which it can be used in other programs as a Java library.

R/Rattle: R (http://www.r-project.org/) and Rattle (R Analytical Tool To Learn Easily) http://rattle.togaware.com/) have been used widely as an open source programming language in research for statistical data analysis. R supports many statistical and data mining techniques, like non-linear modeling, classical statistical tests, time-series analysis, classification clustering, Partial Least Squares Regression (PLSR), random forest and many more. Using R for data mining requires learning the R programming language, an effort that should not be underestimated. However, the R package data mining plugin - Rattle GUI - is available (Williams, 2011) and provides the user with a user-friendly introduction to the methods of data mining. Rattle uses a graphical user interface for an easy use of data mining algorithms based on the R-framework. Rattle GUI activates data mining functions in R that have been implemented.

WEKA: WEKA (Waikato Environment for Knowledge Analysis) (http://www.cs.waikato.ac.nz/ml/weka/index.html) is a Java library with a large selection of sophisticated machine learning algorithms. WEKA can be used in custom Java applications or via the integrated graphical user interface. WEKA contains tools for data preprocessing, classification, regression, clustering, association rules and visualization. WEKA is probably the oldest and most

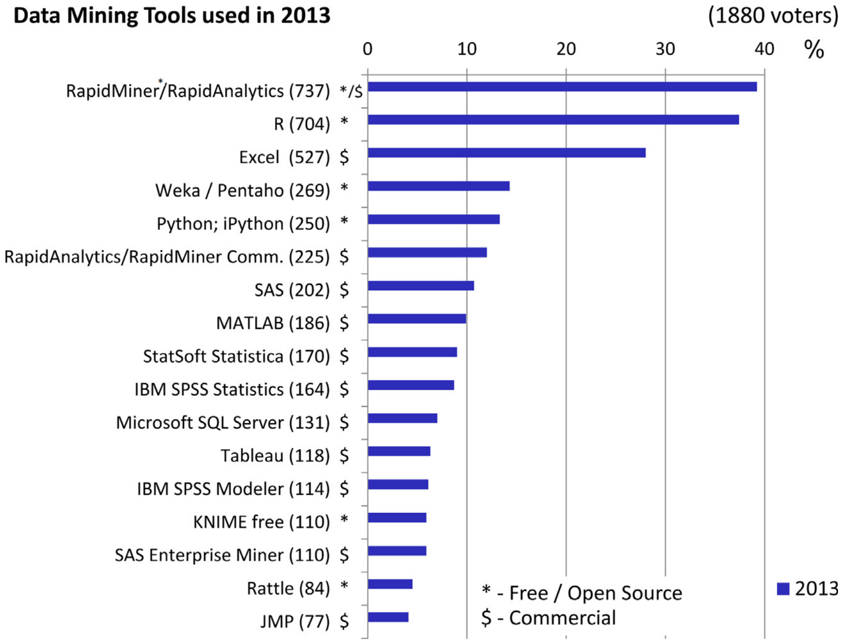

Fig. 3. Report of the KDnuggets Annual Software Poll: Using data mining tools in 2013xps13#(1880 voters, KDnuggets Annual Software Poll, 2013).

successful open source data mining library and software which was later integrated as libraries in RapidMiner and R. Since its GUI is rather rudimentary and offers low user comfort, the usage of WEKA tools has fallen sharply in recent years in favor of RapidMiner and R.

KNIME: (Konstanz Information Miner, http://www.knime.org). Similar to the above, KNIME was developed at the University of Konstanz in 2004 as a machine learning tool. Written in Java, KNIME aimed to provide high operational modularity, while at the same time being simple to upgrade. KNIME provides open source data integration under the Gnu Public License (GPL). KNIME has a very clear structure due to its modularity and is therefore easy to use without any prior knowledge of computer programming. Its ease of use is based on "streamed data processing" that is similar to the knowledge flow mode of WEKA and RapidMiner. All functions such as input, visualization, data manipulation or mining algorithms are packed into so-called "knots". These knots can be moved at will using drag and drop and connected by so-called pipes. In the basic version KNIME only offers basic data mining functions such as data import and export, database access, data manipulation and visualization as well as decision trees, association rules, regression, meta-learning or Bayes clustering. These

Table 1
Comparison of open-source data mining tools.

Data mining tool	WEKA	KNIME	R/Rattle	RapidMiner
Platform	independency	Java (platform-independent)	Java, (platform-independent) based on Eclipse, only executable as a plugin for Eclipse	Platform independent script
Java (platform-independent)				
Target group	learners, advanced users, professional users	learners, advanced users	advanced users, professional users	learners, advanced users, professional users, enterprise users
User interface	GUI	Console display with four different displays. Several projects can be opened at the same time	Conditionally possible in GUI mode if GUI is self-constructed; server mode (command line) and access from Java API	Client/Server architecture: GUI for process design, server execution and access from Java API or Web services, extremely user friendly. Color differentiation for operators and results
Efficiency	Manual setup of threads running simultaneously	Manual setup of threads running simultaneously	Simultaneous processing option	Parallel execution of sub-processes with the help of the parallel processing extension; RapidMiner Server also provides large-scale computation server functionality and parallel execution of processes.
Functions	ca. 500 functions	ca. 100 functions	Data mining functions only extensive with the extension plugin R/Rattle	Greatest range of functions > 500,By incorporating additional extensions the range of functions can be extended tremendously–e.g., through WEKA and R-data mining functions
Interface	Four different modes are available. These modes have their advantages and disadvantages but for every use there is a suitable mode.	A very concise and user-friendly interface, which has been designed accordingly and is easy to learn to use.The interface can easily be configured by the user.	Only supports command line interactions. Several GUI implementations are available for different uses)	A very concise and user-friendly interface, which has been designed accordingly and is very easy to learn to use.The interface can easily be configured by the user.
Supported input formats	ARFF, CSV, C4.5, BSI, local files, from URLs or directly from JDBC-databases	ARFF, CSV, PMML, local files, from URLs or directly from JDBC-databases	A wide range of different input formats are supported, greatly depending on the packages installed	Files: CSV, Excel, XML, Access, AML, ARFF, XRFF, SPSS, SASDatabases: directly from JDBC-databases like MySQL, Postgres, MS-SQL Server, Oracle, IBM DB2, etc.Web: files from URLs, Web services, RSS feeds and many more
Export/ import of models	NA	Models can be exported as zip-files. Only those models that were exported from KNIME can be imported again.	NA	Models can be exported in different file formats, incl. different bitmap formats (bmp,jpg etc.), pdf, as well as vector-formats such as PostScript files, .raw-files etc., Models can be imported as .raw-files and XML-files. Support for PMML.
Extension options	Implementation of one's own algorithms only	Easy and rapid extension of many functions. Both the integration of own as well as readily available third party knots are possible. Such readily available packages contain among others new IO-formats, WEKA-algorithms, R-functions and more.	A huge number of modules (>3000) are available to extend the functions of R	Extensive extensions are already available that tremendously improve its functionality e.g., R-extension, WEKA extension, text extension, web mining extension, image mining extension, linked-open data extension, anomaly detection extension and many more are available on dedicated marketplace.Own extensions can be constructed to extend the range of functions and new functions can be created easily and directly within RapidMiner.
Special features	Experimenter enables one to verify different algorithms with one another, which can be carried out simultaneously on several computers	NA	Large circle of users, particularly well used in research, detailed documentation on all functions	Currently the market leader with the largest circle of users; extensive documentation but as a result of the enormous range of functions not always easy to find

basic functions can be extended in KNIME using pre-assembled plugins from WEKA or R. KNIME is also a plug-in of the popular integrated development environment (IDE) “Eclipse”. A simple comparison of open-source data-mining tools can be seen in Table 1. For more comprehensive comparisons of open-source as well as commercial data mining tools see Hirudkar and Sherekar, 2013.

4. Data mining and linked open data for environmental research

A Web of Science report provides an overview of the use of data mining techniques over the last 20 years (Fig. 4). In 2012, roughly 10% of all scientific papers used data mining methods. An analysis of data mining usage in research areas reveals that data mining techniques are primarily used in computer science and engineering. To date, environmental sciences account for only 1.8% of data mining usage.

Data mining methods are used in many areas such as environmental research and natural hazards. Different applications require specific data mining like spatial data mining for natural hazards, text and web mining in social research and image mining in medical research (Table 2).

Another important area for data mining is earth observation based on remote sensing. Remote sensing techniques produce an enormous amount of spatial, multi-spectral and multi-dimensional earth observation (EO) data. Such data volumes and their inherent complexity necessitate the automatic recognition of spatial patterns, knowledge and insight.

Hyperspectral remote sensing data with a spatial resolution of 0.5 to 3 m and a spectral resolution of 300–400 spectral bands in a spectral range of 400–2500 nm are the most complex EO data available today. Feilhauer et al. (2014), as well as Lausch et al. (2013, 2013a,b,c) have been able to use such hyperspectral data recorded from aircraft to establish numerous process-pattern interactions between soil characteristics and vegetation in environmental research (see Table 2). Fig. 5 shows an example for spatial data mining based on airborne hyperspectral data–AISA-EAGLE/HAWK (Airborne Imaging Spectrometer for Applications) at the TERENO–Schäfertal test site. The goal of these hyperspectral measurements is to identify important soil characteristics like organic matter, texture parameters, clay content or moisture structures and important biochemical–biophysical vegetation characteristics (chlorophyll, water- cellulose content). The vegetation parameters are the proxy or functions of soil components and moisture patterns. In this way, important soil process-vegetation patterns and interactions are modeled based on data mining methods (Lausch et al., 2013).

The advantages of such spatial data mining techniques arise from their ability to capture and determine unknown, interesting and significant patterns and knowledge in spatial as well as temporal high resolution EO data. The highly multi-dimensional 300–400 spectral bands as well as multiple EO datasets over the year necessitate partial and fully automated data-mining techniques for research. Such knowledge about patterns and processes is an absolute necessity for subsequent modelling and projections about environmental issues Table 3.

The greatest disadvantage of data mining compared to LOD, (see Section 4) is that with data-mining methods, patterns and knowledge can only be extracted from closed or “silo” data sets, where all of the data required for the analysis are associated through known relationships as established in relational databases. Consequently, it is the job of the analyst to couple all of the necessary data and information that could potentially influence patterns and processes. Should the analyst make an incorrect decision in the process of selecting suitable variables for the model, this cannot be taken into account during the modelling process. Hence, it is possible that new or unknown patterns and knowledge are not recorded due to incorrect decisions that have been made previously.

5. Demands of linking future research and data mining

All of the previously discussed applications of data mining in different research areas show that the usage of data mining techniques is mostly tailored to certain areas. For example, social sciences primarily use text and web mining for analyzing surveys, markets or socio ecological patterns whereas medical research primarily uses methods of image mining to identify and predict cancer or classical data mining to identify relationships between the demographics of patients with critical illnesses. Environmental research extracts patterns based on multidimensional and complex spatio-temporal remote sensing data for the automated discovery of spatial knowledge. This mainly area-based data mining approach will, however, no longer be expedient.

In order to gain an understanding of ecological patterns and their importance and effects on the driving process in environmental research, economic, social, environmental and ecological indicators have to be included in modeling and projections in addition to the ecological patterns.

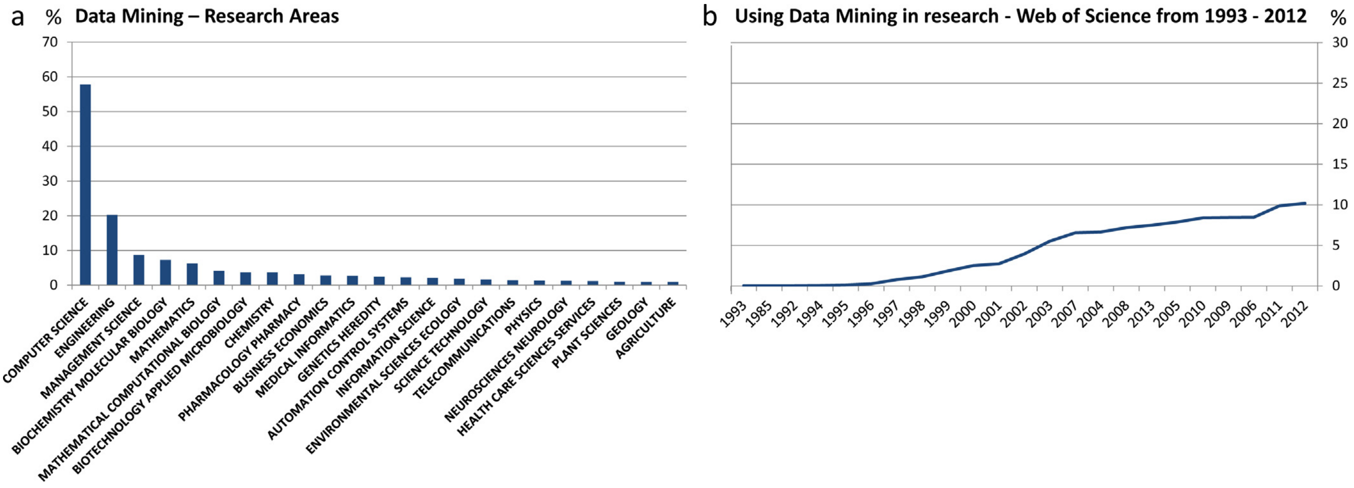

Fig. 4. Citation report in the Web of Science with the topics - Data mining–(a) Using Data Mining in Research area, (b) Data Mining in Research–Web of Science from 1993xps14#2012.

Table 2
Overview of conventional data mining applications in different research areas.

Research	Types of data mining	Area of application	References
Environmental research	Spatial data mining	Analysis of data for earth observation, Geoinformation Data	Lausch et al., (2013, 2013a,b,c); Wu et al., 2013; Feilhauer et al., 2014
Natural hazards	Spatial data mining Web Mining	Flood damage assessment,Inventory of geohazards at national scale	Merz et al., 2013; Battistini et al., 2013
Social research	Web mining text mining	Social network analysis, life science	Murthy et al., 2013; Touw et al., 2013
Education	Data mining	Analysis of factors for learning success	Baker et al., 2008; Bhise et al., 2013
Medicine	Spatial data mining image mining data mining	Diagnostic, tumor diagnostic, mammography, risk assessment in medicine, correlating demographics of patients with critical illnesses,	Burget et al., 2011; Touw et al., 2013; Mohanty et al., 2013
Biology	Data mining spatial data mining text mining	Biotechnology, protein-genom-DNA analysis, spatial epidemiology, plant research	Brady and Provart, 2009; Alonso-Peral et al., 2012; Landeghem et al., 2013
Chemistry, pharmaceutical research	Data mining	Analysis chemical reactions, material analysis	Hautier et al., 2012; Curtarolo et al., 2013
Economic research	Data mining	Finding predictive information, risk assessment, marketing, customer ratings, quality assurance, financial-data analysis, financial control, e-commerce, financial-data analysis, business analysis	Maity and Chatterjee, 2012; Ledolter, 2013; Jain and Singh, 2013
Medien Research	Mining of image, video, music	Image analysis, indexing of image, video, music	Yuan et al., 2011; Saravanan and Srinivasan, 2013; Kumar and Saravanan, 2013
Computer hardware and software science	Data mining	Predicting disk-failures and potential security violations, software engineering	Vintrou, 2013; Ikonomovska et al., 2011

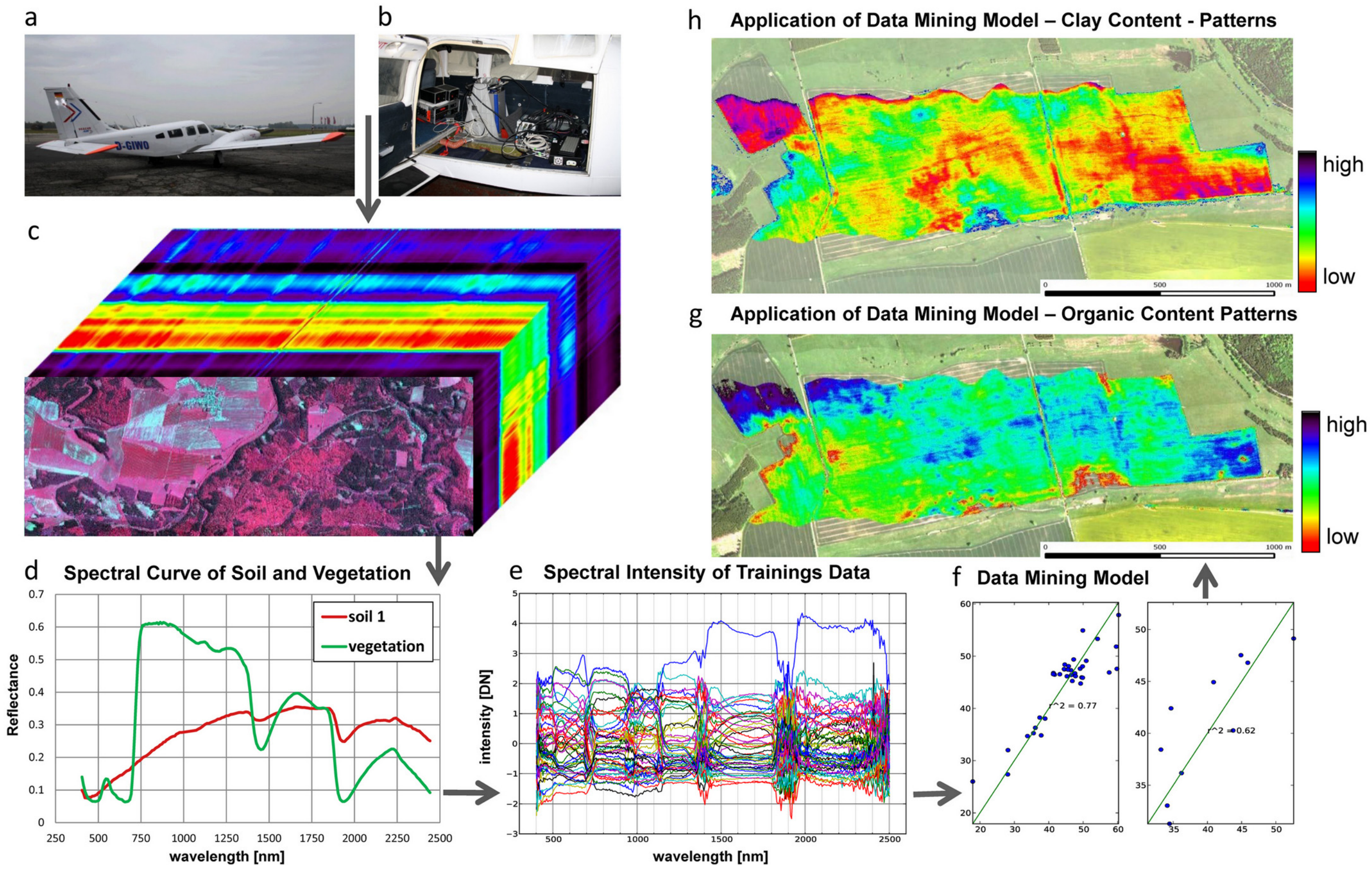

Fig. 5. (a–b) Airborne hyperspectral AISA-Eagle/HAWK remote sensor mounted on Piper, (c) CIR-image from hyperspectral sensors of the AISA-EAGLE/HAWK (AISA-DUAL) 400–2,500 nm with data cube, 367 spectral bands with 2 m recorded ground resolution, date of recording Mai 2012 with a Piper, Region Schäfertal - Bode Catchment, (d) Spectral curve of ground truth sampling points for soil and vegetation in the test site. (e) Spectral intensity curves of imaging hyperspectral data, (f) Data Mining Model, (g) Application of the best data mining model on airborne hyperspectral image data for quantification and recognition organic content patterns and (h) Pattern of Clay content.

Table 3
Requirements of data mining tools, techniques and platforms for future interdisciplinary research and possible solutions.

Requirements	Possible solutions
Linking spatial reference, environmental, social, economic research among others in complex data mining surveys	Linked open data–LOD[a]
Linking various data mining types (text, music, video, image, spatial-temporal data mining) with the goal of a joint analysis	Linked open data–LOD[a]
Providing freely accessible data from the various areas of environmental research to work together on key research areas	Open data initiative[b]
Providing open data bases containing semantic data that are accessible online	Open databank initiative - freebase[c]
Open source software for simple yet complex data management	Talend open studio[d]
Using the Internet as a computing platform	Hadoop®[e]
Simple handling and use of data mining tools, techniques, and libraries that does not require a big time investment and in-depth programming knowledge	RapidMiner[f] RapidMiner®/R/ Weka-couppling
Development and provision of data mining platforms, which enable joint project work "under one roof" to generate models, hypotheses and forecasts based on data mining and LOD	RapidAnalytics®[g]

[a] Linked Open Data–LOD http://linkeddata.org/.
[b] Open Data Initiative http://de.wikipedia.org/wiki/Open_Data.
[c] Open Databanks–Freebase http://freebase.com.
[d] Talend Open Studio http://de.talend.com/products/talend-open-studio.
[e] Hadoop® http://hadoop.apache.org/.
[f] RapidMiner® http://rapidminer.com/.
[g] RapidAnalytics® http://www.e-lico.eu/rapidanalytics.html.

Therefore environmental research will face several challenges now and in the future. These are:

- Dealing with the enormous increase in data volume and complexity from all disciplines.
- A very interdisciplinary nature and interactions of various disciplines like sociology, economics, law, ecology, toxicology and others.
- As a result of the impact of economic, social, environmental and planning decisions, complex data that is currently available has to be brought into perspective with its historical development and historical change.

6. Linked open data–A new data mining approach in research

LOD is a recent development whereby all data sets that are freely available on the Internet are interrelated by way of semantic nets and techniques (Dengel, 2012) in an effort to publish these in a machine readable way and to facilitate analyzing them in human-understandable form (Heath and Bizer, 2011).

6.1. Semantic methods and networks

Semantic networks are a visual form of knowledge representation (Dengel, 2012). With the help of semantic networks, statements about objects using nodes and the relationships between them can be graphically displayed. This helps the meaning of an object (its definition) to become ascertainable through its associative relationship with other objects (definitions, data). Associative relationships can express the relationship between two completely independent objects (definitions) that do not have to have any connection with each other, but can be loosely associated under certain conditions. In Fig. 6 a simple semantic network is shown using the example of the sentence "Uta makes coffee for Lutz". The subject and object stand for nodes, whereas the predicates and verbs are represented by edges.

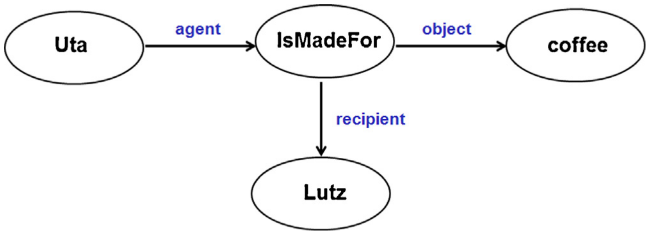

Fig. 6. Semantic network for the sentence "Ute makes coffee for Lutz".

To create LOD and to carry out LOD queries, clearly defined rules need to be established for the associative relationships between the objects. The rules for the statements - Oskar "is a" man and - Antje "is a" woman would for example be as follows (adapted after Dengler, 2012):

- if B "is the child of" A ***and*** A "is a" man, then A "is the father of" B
- if B "is the child of" A ***and*** A "is a" woman, then A "is the mother of" B

A, B = variables for the objects to be investigated

From the above it is possible to make the following two statements: (1) All men who have children are fathers, and (2) all women who have children are mothers. Fig. 7 shows a more complex semantic network to describe the relationships between the family members in Oscar's family. Semantic networks can be static or change over time. At the same time, semantic networks can expand very rapidly, in which case the derivation of knowledge is required using data mining. Further information on semantic networks can be found in Dengel (2012) and Reichenberger (2012).

6.2. Concept of LOD

The concept of LOD is based on the idea of linking publicly available data "silos" on the internet by means of semantic methods and networks. By linking data, all of the data objects (e.g., man, woman) become related to each another. By determining a number of rules about these relationships, such inter-linked data can be "understood" by machines and algorithms, which enables global data mining approaches and the discovery of truly new associations, patterns and knowledge.

LOD is based on the Resource Description Framework (RDF) data model, which formulates syntax and rules about data and resources as well as their location on the internet. The RDF-model resembles classic methods for concepts such as the Entity-Relationship-Model (ERM) or the Unified Modelling Language (UML) class diagram. RDF models have a formal semantic that is based on digraphs. The RDF syntax extends the Extensible Hyper Text Markup Language (XHTML) of the web by semantic annotation in support of linking up data. Along with RDF there are also other standards such as the Web Ontology Language

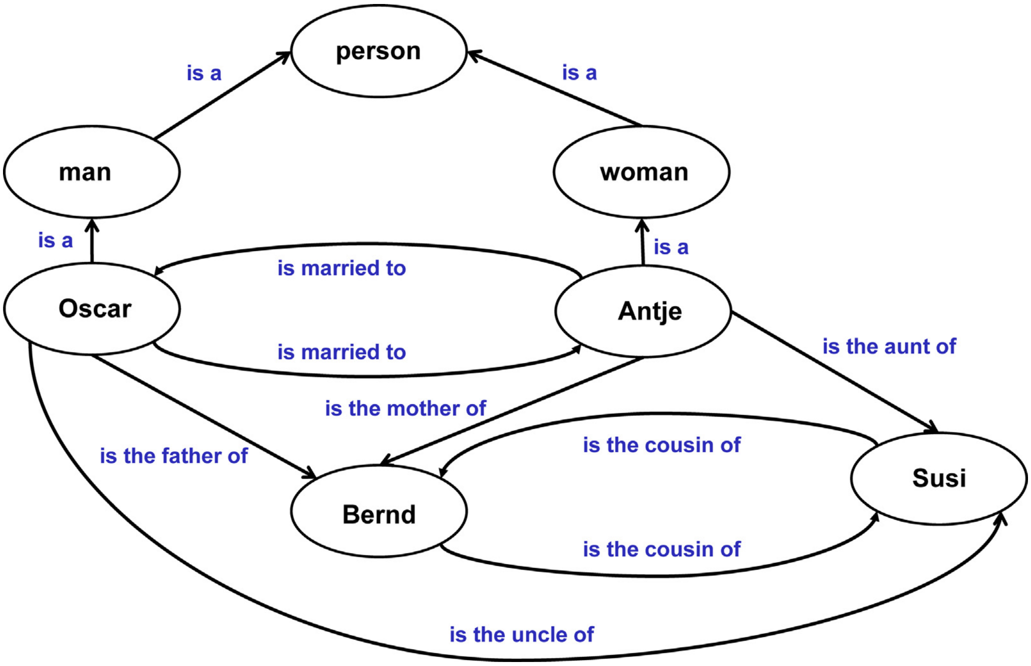

Fig. 7. Semantic network to describe the relationships of Oscar's family members with one another.

(OWL) that enable more complex semantic annotations. More detailed literature on the subject can be found in Dengel (2012) as well as under http://www.w3.org.

The LOD approach is not about expanding linkage of documents via hyperlinks–something which has been common on the Internet - but, rather, about the "linking of all publically accessible data" together (Heath and Bizer, 2011). Publically accessible data includes many areas of interest, such as weather, events, politics, business, news, geographic spatial data, videos, pictures, images and much more. In this context, Berners-Lee (2006, 2009) refers to the LOD approach as the next generation of the web.

Implementation of LOD approaches requires adherence to the four basic components as formulated by Berners-Lee (2006) and Dengel (2012):

1. Use Uniform Resource Identifiers (URI) to uniquely identify data.
2. Use the Hypertext Transfer Protocol (HTTP) so that people, web agents and data mining tools can access and refer to data.
3. A URI has to refer to usable information that can be provided with the RDF and queried with the Simple Protocol and RDF Query Language (SPARQL).
4. Links to other RDF resources should be established in support of growing a word wide network of publically available data and allowing for truly inter-disciplinary data mining.

According to Berners-Lee and his call in a TED lecture (Berners-Lee, 2009) for "raw data now!" more and more people, institutions, companies and research centers are putting their data on the web in a machine readable form as per the LOD approach thereby making it available to the general public for complex and interdisciplinary data analysis. In this way, the linked data on the web results in a worldwide LOD network which according to Bizer et al. (2009) is also referred to as the "linked open data cloud". Google, for instance, uses LOD to improve search functions. DBpedia is an LOD version of the open and free online encyclopedia Wikipedia. In August 2011, the LOD cloud numbered roughly 31.6 billion RDF Tripel, linked over 389 million RDF links (Dengel, 2012).

7. LOD in medical research

The semantic search in medical images was first successfully implemented in the research project THESEUS MEDICO (Lavrac and Novak, 2013). The objective of LOD was, on the one hand, to simplify access to online information and data and, on the other hand, to network data and generate new knowledge by linking complex and heterogeneous data.

Imaging techniques like Computer Tomography (CT) and Magnetic Resonance Tomography (MRT) are key procedures in medical diagnostics and treatment and aim to recognize and diagnose sicknesses early on and treat them correctly. The symptoms identified in CT and MRT are often difficult to diagnose as a result of the tremendous complexity stemming from various causes. A targeted treatment is, for this reason, not always possible because the causes and interactions of symptoms cannot be clearly ascertained. By linking complex and not just stand alone specialist information and data, the linked open data approach offers tremendous new potential for future research.

Fig. 8 shows a conceptual LOD model of the semantic linking of open data in medical research using the example of disease diagnosis through computer tomography. In the LOD approach, heterogeneous and complex data like texts, images or lab data are semantically linked with each other.

The advantage of the LOD concept here is that while there is a "subject core" - in this case the CT/MRT symptom - a number of links to other disciplines like sociology, toxicology, biochemistry can be made as a result of linked open data, a fact which makes it possible to identify patterns in interactions between symptoms and complex causes.

8. LOD in environmental research

Coupling of Database–EUROSTAT: Paulheim et al. (2013) developed the first linked open data extension for the open source tool RapidMiner. This example shows how to read and analyze data from the linked open data source, Eurostat (Paulheim and Fürnkranz, 2012). The corresponding RapidMiner workflow

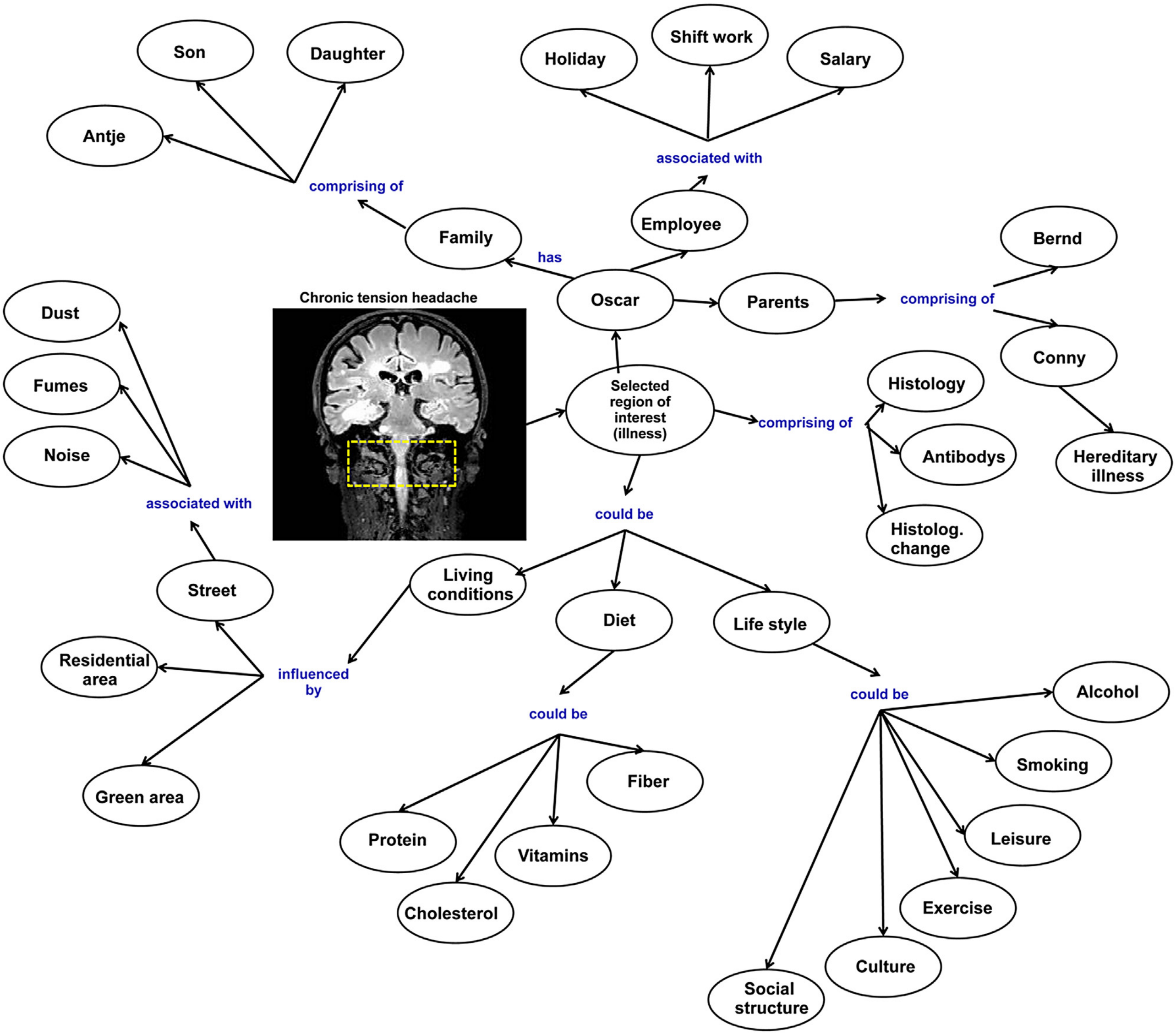

Fig. 8. Conceptual LOD Modell of semantic link of open data in medical research using the example of disease diagnosis through magnetic resonance tomography (MRT). The clinical MRT picture (yellow box) shows the classical symptoms of chronic tension headache (example). This illness is effect by different causes. The importance of causes can be analysed with data mining of linked open data approach.

(Fig. 9) implements data mining techniques of Eurostat data for analyzing patterns based on the semantic linked open data. Based on a classification tree as data mining method the semantically linked LOD data from Eurostat is researched with regard to which factors explain high alcohol consumption in European countries.

Remote Sensing Research: The growing volume of Earth Observation (EO) data (through global monitoring, long-term archives–(European Remote Sensing Satellite - ERS, Environmental Satellite - Envisat) and the associated growth in diversity and complexity of additional data, methods and models requires the creation of data networks and Thematic Exploration Platforms (TEP, ESA, 2014). The principal idea of TEP is “to move processing to the data, rather than the data to the users” (ESA, 2014). First successful applications of TEP involve the Geo-spatial Processing on Demand (URL, 2014), environment. The “Earth Observation Exploitation Platform (EOXP, Bitto et al., 2014) prototype illustrates another possible implementation of TEP. TEP in combination with LOD (Balazinska et al., 2007) will be the foundation for a world-wide inter-connected data archive, which will serve as a base for truly global and interdisciplinary data mining of remotely sensed and associated (ground truth and meta) data. A world-wide network of remotely sensed data will facilitate rapid data access and enable a fully automated data processing framework. New tools and cost-effective data mining approaches will make this potential available for researchers of all fields, independent of their expertise in computer programming and mathematics (ESA, 2014). More examples of applications of semantic networks and LOD are shown in Table 4.

8.1. Advantages and tools of the LOD approach

The advantage of the LOD approach relies on the possibilities of retrieving and subsequently analyzing data that have been related through associative links. The data range, the depth of branching and thus the range of topics that are linked with one another depend on the opening, provisioning and linking of data all of which are constantly growing on the Internet through RDF semantic. Through LOD it will therefore become increasingly possible to discover completely new relationships, patterns, insights and knowledge.

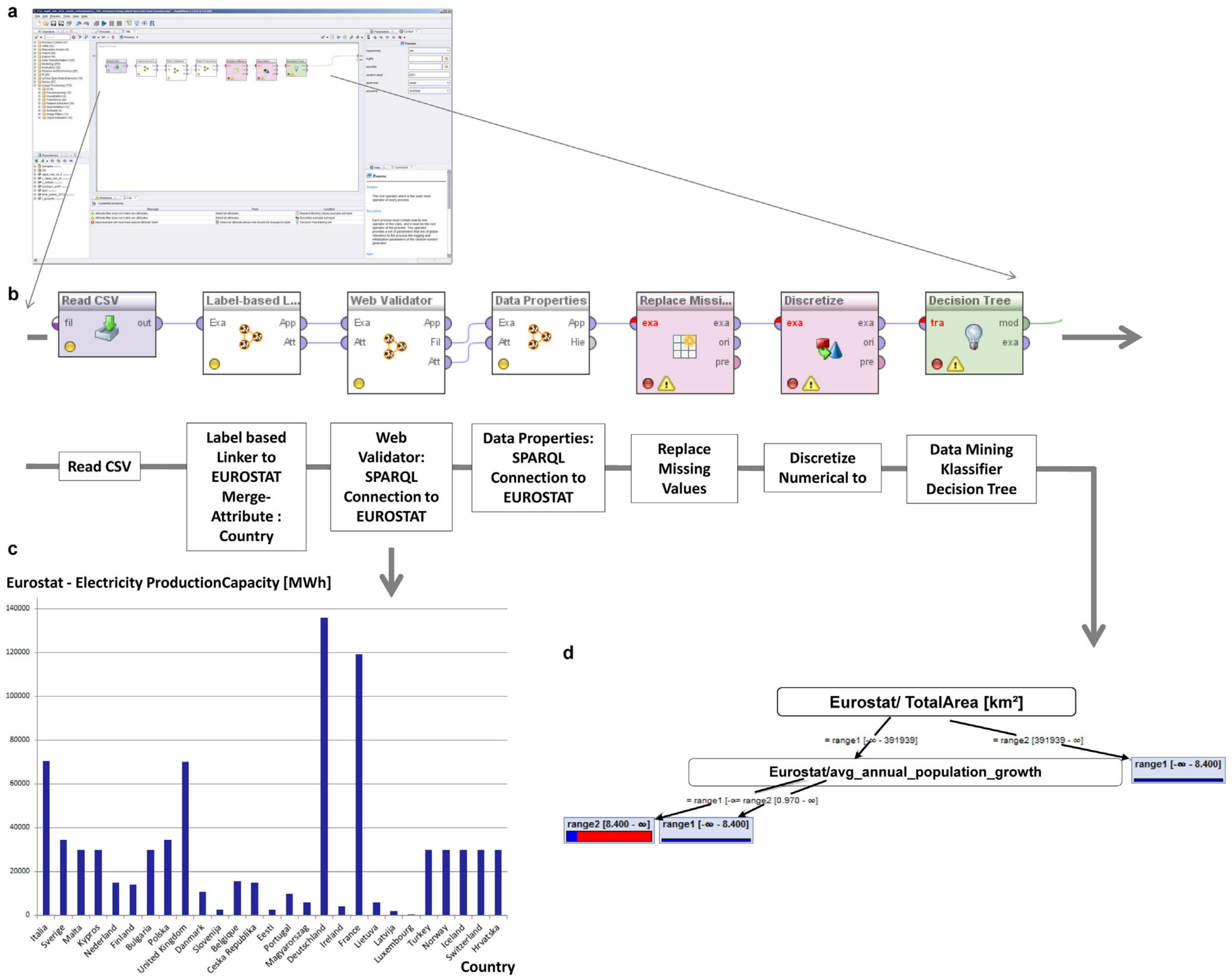

Fig. 9. Application for linking open data of EUROSTAT based on the open source data mining tool RapidMiner. (a, b) Process procedure of data mining in RapidMiner, (c) Singel results of LOD data mining analysis from Eurostat based on LOD, Electricity Production Capacity [MWh]/Country, (d) Data mining application–Classification tree of EUROSTAT data to explain the high alcohol consumption in different countries (modified after Paulheim and Fürnkranz, 2012).

Table 4
Overview of LOD applications in different research areas.

Research	Area of application–semantic Web/LOD	References
Environmental research	Land cover class extraction in Geographic Object-Based Image Analysis (GEOBIA) using environmental, spatial and temporal ontology	Aryal et al., 2014
	Development of large-scale worldwide sensor networks and "Earth Observation Exploitation Platform (EOXP) based on semantic, ontology and Linked Open Data approaches	Hart and Martinez, 2006; Balazinska et al., 2007; Henson et al., 2009; Bitto et al., 2014
	Earth,Life and Semantic Web (ELSEWeb): coupling of earth observation data sources with biological information and ground truth data for biodiversity forecasting,	Del Rio et al., 2014
	Virtual Observatory Infrastructure for Earth Observation Data (TELEIOS), mapping, analyzing and prediction of climate change, real time fire monitoring based on Semantic, Ontology and LOD, environmental monitoringhttp://www.earthobservatory.eu/	Koubarakis et al., 2012; Boic et al., 2014; Stocker and Rönkkö, 2014
	Earth observation data classification, managing and environmental modelling of complex environmental data	Villa et al., 2009; Lockers et al., 2014; Nieland et al., 2014
Economy and social networks	Crating an Open Government Data (OGD) Platform for coupling governments and public authorities data like Eurostat, World Bank, OECD, CIA's World and Facebook.	Kalampokis et al., 2013
Disaster mitigation	Disaster mitigation and preparedness using linked open data	Thushari et al., 2012
Data management	Managing very diverse data for complex, interdisciplinary science	Parsons et al., 2011
Network Learning	Research on network learning	Goodyear et al., 2004

Table 5
Commercial and open source tools for storing, publishing and managing LOD.

Tools for LOD	Weblinks
RapidMiner, LOD extension	http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension/. http://marketplace.rapid-i.com/UpdateServer/face/product_details.xhtml?productId=rmx_lod.
4Store	http://neologism.deri.ie/.
Mulgara	http://www.mulgara.org/.
Apache Jena framework	http://incubator/apache.org/jena/index.html/.
Sesame	http://www.openrdf.org/.
Oracle 11 g database	http://www.oracle.com/technetwork/database/options/semantic-tech/whatsnew/index.html/.
Allegro graph	http://www.franz.com/agraph/allegrograph/.
Big data	http://www.systap.com/bigdata.com/.
Open link virtuosa	http://www.virtuosa.openlinksw.com/.
Hyperdata browser–disco	http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/.
Open link data explorer	http://ode/openlinksw.com/.
Linked Geo data	http://browser.linkedgeodata.org/.
Semantic web client library	http://dataviewer.zitgist.com/.

Another advantage of LOD relies on the fact that the data are very interdisciplinary and there is a diversity of possible data formats that can be integrated into a LOD analyses. Because the coupling of LOD data is standardized by means of RDF rules and syntax, no further steps are required for data transformation or data integration.

The development of tools that support LOD queries is constantly growing. Table 5 contains commercially available and open source tools for storing, publishing and managing open linked data.

9. Conclusion

The rapid development of information technology in the 21st Century is associated with an enormous increase in data volumes and data complexity. The insights gained from such data complexity and data volumes are not holding up to our expectations as modern data analyzing techniques are not yet widely known and used.

Data mining offers one promising solution to this situation with numerous approaches for knowledge discovery in multi-dimensional and complex data. Unfortunately, data-mining techniques are currently exclusively used in analytical research areas such as computer science, physics or mathematics and often require extensive programming expertise.

The objective of this paper was therefore to make researchers with little or no experience in computer programming aware of the significance and tremendous potential of data mining in a demonstrative manner. To achieve this, data mining processes, types and methods were portrayed in a comprehensible manner and important references were provided to existing open source data mining tools, which are characterized by their simple and intuitive use in support of complex data analysis.

Good, simple and easy-to-learn open source data mining tools are RapidMiner and KNIME, which enable very complex data-mining processes to be generated without any previous computer programming knowledge with the help of intuitive graphical user interfaces. Many data mining functions can be added to those tools by means of plugins and connectors to WEKA and R/Rattle are also available.

LOD is a novel approach that can help us to discover new insights and knowledge from truly multi-disciplinary and interlinked data sets publically available on the internet. This paper comprehensively explains the concept of LOD. Reference is made to the basic elements of semantic networks, LOD, RDF, SPARQL and OWL and supported by examples whenever possible. In comparison with traditional data mining techniques, reference is also made to the advantages of LOD as well as existing commercial and open-source LOD tools.

The LOD approach is a new concept for data integration with a huge potential for analyzing interdisciplinary, complex and distributed data sets. According to Berners-Lee (2006, 2009) LOD will be the next generation of the Internet by associating publicly accessible data in a standardized way. One huge advantage of LOD compared to traditional data mining is that the analyst does not have to link associated data prior to analysis, but that machine learning algorithms will be able to analyze an open set of interdisciplinary databases.

The LOD approach will facilitate to find novel insights from and novel associations between interdisciplinary and complex data sets and based on such knowledge we will be able to solve complex problems and make better predictions.

References

Alonso-Peral, M.M., Sun, C., Millar, A.A., 2012. MicroRNA159 can act as a switch or tuning microRNA independently of its abundance in Arabidopsis. PLoS ONE 7, e34751.

Alpha, W., 2013. http://www.wolframalpha.com/about.html (Date: 1.12.2013).

J. Aryal A. Morshed R. Dutta 2014. Land cover class extraction in GEOBIA using environmental spatial temporal ontology Proceeding of the South-Eastern European Journal of Earth Observation and Geomatics Special Issue, 21–24 May 2014, Thessaloniki, Greece, ISSN 2241 2014 429–434.

Balazinska, M., Deshpande, A., Franklin, M., Gibbson, P.B., Nath, J.G., Hansen, M., Liebhold, M., Szalay, A., Tao, V., 2007. Data Management in the Worldwilde Sensor Web. IEEE Comput. Soc. 1536–1568.

Battistini, A., Segani, S., Manzo, G., Catani, F., Casagli, N., 2013. Web data mining for automatic inventory of geohazards at national scale. Appl. Geogr. 43, 147–158.

R. Baker T. Barnes J.E. Beck. Educational Data Mining. 1st. International Conference on Educational Data Mining, Proceedings. Montreal, Quebec, Canada.

Begum, S.H., 2013. Data Mining Tools and Trends–An Overview. International. J. Emerg. Res. Manage. Technol. 2278–9359, 6–12.

Bensberg, F., Weiß, T., 1999. Web Log Mining als Marktforschungsinstrument für das World Wide Web. Wirtschaftsinformatik 41, 426–432.

Berners-Lee T., 2006. Linked Data. W3C Design Issue. URL http://www.w3 org/DesignIssues/LindedData.html, (Date: 20.07.2014).

Berners-Lee T., 2009. The Next Web. URL http://www.ted.com./talks/tim_berners_lee_on_the_next_web.html, (Date: 20.07.2014).

Bhise, R.B., Thorat, S.S., Supekar, A.K., 2013. Importance of Data Mining in Higher Education System 6, 18–21.

Bitto, F., Castracane, P., Graziano, A., Lapaola, M., Farres, J., Zelli, C., 2014. The Earth Observation Exploitation Platform (EOXP) Prototype, Technical Note, March 2014. https://earth.esa.int/documents/10174/1157689/EOXP-ACS-ESA-IIM (Date: 20.07.2014).

Bizer, C., Heath, T., Berners-Lee, T., 2009. Linded Data–The Story So Far. Int. J. Seman. Web and Info. Syst. 5, 1–22.

Boic, B., Peters-Anders, J., Schimak, G., 2014. Ontology Mapping in Semantic Time Series Processing and Climate Change Prediction. Proceedings of International Environmental Modelling and Software Society (iEMSs), San Diego, http://www.iemss.org/society/index.php/iemss-2014-proceedings.

Brady, S.M., Provart, N.J., 2009. Web-queryable large-scale data sets for hypothesis generation in plant biology. Plant Cell 21, 1034–1051.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Monterey, CA, Wadsworth.

R. Burget V. Uher R. Cervenec Rapid Image Feature Extraktion and Mining Fischer, S Mierswa, I. (Eds.) Proceedings of the 2nd Rapid Miner Community Meeting and Conference, RCOMM Aachen 2011 2011 123–132.

Chu, W.W., 2014. Data Mining and Knowledge Discovery for Big Data. Methodologies, challenge and Opportunities. Springer Verlag Berlin Heidelberg.

Computer, 2013. Statista–Das Statistik-Portal. Das Internet im Jahr 2015. Computer Bild 22, 116-117.

Curtarolo, S., Hart, G.L.W., Nardelli, M.B., Mingo, N., Sanvito, S., Levy, O., 2013. The high-throughput highway to computational materials design. Nat. Mater. 12, 191–201.

Del Rio N., Villanueva-Rosales, N., Pennington, D., Benedict, K., Stewart, A., Grady, C. J., 2014. ELSEWeb Meets SADI: Supporting Data-to-Model Integration for Biodiversity Forecasting. Discovery Informatics: AI Takes a Science-Centered View on Big Data AAAI Technical Report FS-13–01, http://www.aaai.org/ocs/index.php/FSS/FSS13/paper/viewFile/7631/7488.

Dengel, A., 2012., Semantische Technologien. Spektrum Akademischer Verlag.

J. Dörre P. Gerstl R. Seiffert Text Mining. In: Hippner H., Küsters U., Meyer, M., Wilde, K.D. (eds.): Handbuch Data Mining im Marketing. Wiesbaden 2001 2001 465–488.

ESA, 2014. THEMATIC EXPLOITATION PLATFORMS. Official Tender of the ESA, 2014. AO7870, http://emits.sso.esa.int/emits/owa/emits.main, (Date: 20.07.2014).

Fayyad, U., Shapiro, G., Smyth, P., 1996. The KDD process for extracting useful knowledge from volumes of data. Commun. ACM 39, 27–34.

Feilhauer, H., Doktor, D., Lausch, A., Schmidtlein, S., Schulz, G., Stenzel, S., 2014. Mapping Natura 2000 habitats and their local variability with remote sensing. Appl. Veg. Sci doi:http://dx.doi.org/10.1111/avsc.12115.

Goele,S., Chanana, N., 2012. Data Mining Trend In Past, Current And Future. International Journal of Computing & Business Research, In Proc. I-Society 2012. http://www.researchmanuscripts.com/isociety2012/15. pdf, (Date: 20.07.2014).

Goodyear, P., Banks, S., Hodgson, V., & McConnell, D. 2004. Research on networked learning: An overview. In P. Dillenbourg, M., Baker, C., Bereiter, Y. Engeström, G., Fischer, H. U. Hoppe, D. McConnell (Eds.), Advances in Research on Networked Learning,Springer, 1–9.

J.K. Hart K. Martinez Environmental Sensor Netzworks: A revolution in the earth science? Earth-Science Reviews 78 2006 177–191.

Hautier, G., Jain, A., Ong, S.P., 2012. From the computer to the laboratory: materials discovery and design using first-principles calculations. J. Mater. Sci. 47, 7317–7340.

Heath, T., Bizer, C., 2011. Lined Data: Evolving the Web into a Global Data Space. Morgan & Claypool, 1st. Edition. http://linkeddatabook.com/, (Date: 1.12.2013).

C.A. Henson J.K. Pschorr A.P. Sheth K. Thirunarayan SemSOS: Semantic Sensor Observation Service. Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems (CTS 2009), Baltimore, MD May 2009 18–22.

Hirudkar, A.M., Sherekar, S.S., 2013. Comparative Analysis of Data Mining Tools and Techniques for Evaluating of Database Systems. Int. J. Comput. Sci. Appl. 6, 232–237.

Hofmann, M., Klinkenberg, R., 2013. RapidMiner–Data Mining Use Cases and Business Analytics Applications. CRC Press, Taylor & Francies Group.

Ikonomovska, E., Gama, J., Dzeroski, S., 2011. Learning model trees from evolving data streams. Data Mining Knowledge Discov. 23, 128–168.

Jain, R., Singh, D., 2013. Data Mining and Analysis of Economic Data. International of Advandced Research in Computer Science and Software Engineering–International Journal of Advanced Research in Computer Science and Software Engineering 3, 683–688.

Joseph, M.V., Sadath, L., Rajan, V., 2013. Data Mining: A Comparative Study on Various Techniques and Methods. Int. J. Adv. Res. Comput. Sci. Software Eng. 2, 106–113.

Kalampokis, E., Tambouris, E., Tarabanis, K., 2013. Linked Open Government Data Analytics, M.A., Wimmer, M. Janssen, and H.J. Scholl (Eds.): EGOV 2013. LNCS 8074, 99–110. IFIP http://dx.doi.org/10.1007/978-3-642-40358-3_9.

KDnuggets Annual Software Poll, 2013 KDnuggets Annual Software Poll: Using Data Science software in 2013. (http://www.kdnuggets.com/2013/06/kdnuggets-annual-software-poll-rapidminer-r-vie-for-first-place.html, (Date: 01.12.2013).

M. Koubarakis M. Karpathiotakis K. Kyzirakos C. Nikolaou S. Vassos G. Garbis M. Sioutis K. Bereta S. Manegold M.L. Kersten M. Ivanova H. Pirk Y. Zhang C. Kontoes I. Papoutsis T. Herekakis D. Mihail M. Datcu G. Schwarz O. Dumitru D. Espinoza-Molina K. Molch U.D. Giammatteo M. Sagona S. Perelli E. Klien T. Reitz R. Gregor Building virtual earth observatories using ontologies and linked geospatial data Krötzsch, M. Straccia, U., (Eds.) Web Reasoning and Rule Systems, RR 2012, LNCS 7497 2012 Springer–Verlag Berlin Heidelberg, 229–233

Kumar, R.A., Saravanan, D., 2013. Content based image reterival using color histogram. Int. J. Comput. Sci. Inform. Technol. 4, 242–245.

Kreissl, R. 2013. Datenspuren: Komplette Umkehr der Beweislast. New Scientist, http://irissproject.eu/?p=325, (Date: 22. Februar 2013).

Landeghem, S.V., De Bodt, S., Drebert, Z., Inzé, D., Van de Peer, Y., 2013 The Potential of Text Mining in Data Integration and Network Biology for Plant Research: A Case Study on Arabidopsis Plant Cell 25, 794–807.

Lausch, A., Heurich, M., Gordalla, D., Dobner, H.-J., Gwillym-Margianto, S., Salbach, C., 2013. Forecasting potential bark beetle outbreaks based on spruce forest vitality using hyperspectral remote-sensing techniques at different scales. Forestest Ecol. Manage. 308, 76–89.

Lausch, A., Pause, M., Merbach, I., Zacharias, S., Doktor, D., Volk, M., Seppelt, R., 2013a. A new multi-scale approach for monitoring vegetation using remote sensing-based indicators in laboratory, field and landscape. Environmental Monitoring Assess. 185, 1215–1235.

Lausch, A., Pause, M., Schmidt, A., Salbach, C., Gwillym Margianto, S., Mehrbach, I., 2013b. Temporal imaging hyperspectral monitoring of chlorophyll: LAI and water content of barley during a growing season. Can. J. Remote Sensing 39, 191–207.

Lausch, A., Zacharias, S., Dierke, C., Pause, M., Kühn, I., Doktor, D., Dietrich, P., Werban, U., 2013c. Analysis of vegetation and soil pattern using hyperspectral remote sensing, EMI and Gamma ray measurements. Vadose Zone J doi:http://dx.doi.org/10.2136/vzj2012.0217.

Lavrac, N., Novak, P.K., 2013. Relational and Semantic Data Mining for Biomedical Research. Informatica 37, 35–39.

Ledolter, J., 2013. Data Mining and Business Analytics with R., Wiley & Sons. New Yersey, Hoboken.

Maity, B., Chatterjee, B., 2012. Forecasting GDP Growth Rates of India.: An Empirical Study. Int. J. Econom. Manage. Sci. 1 (9), 52–58.

Merz, B., Kreibich, H., Lall, U., 2013. Multi-variate flood damage assessment: a tree-based data-mining approach. Natural Hazards Earth Syst. Sci. 13, 53–64.

Meyer, M. Lüling, M., 2003. Data Mining in Forschung und Lehre in Deutschland 4, Schriften zur Empirischen Forschung und Qualitativen Unternehmensplanung 15, 2003. Ludwig-Maximilians-Universität München. http://www.imm.bwl.uni-muenchen.de/forschung/schriftenefo/ap_efoplan_15 pdf. (Date: 1.12.2013).

Mohanty, A.K., Senepati, M.R., Lenka, S.K., 2013. A novel image mining technique for classification of mammograms using hybrid feature selection. Neural Comput. Appli doi:http://dx.doi.org/10.1007/s00521-012-0881-x.

Murthy, D., Gross, A., Takata, A., Bond, S., 2013. Evaluation and Development of Data Mining Tools for Scial Netwirk Analysis. Özyer, T. (eds.) Mining Social Networks and Security Informatics Lecture Notes in Social Networks. Springer Science +Business Media Dordrecht, DOI 10.1007/978-94-007-6359-3_10. 183–202.

Nieland, S., Moran, N., Kleinschmit, B., Förster, M., 2014. Using semantic-based spatial reclassification for interoperable data management in Natura 2000 monitoring. Proceedings of International Environmental Modelling and Software Society (iEMSs), San Diegeo, http://www.iemss.org/society/index.php/iemss-2014-proceedings.

Paulheim, H., Mitichkin, E., Ristoski, P., Bizer, C., 2013. RapidMiner Linked Open Data Extension. Manual Version 1.2,11/29/13. University of Mannheim, Data and Web Science Group, http://dws.informatik.uni-mannheim.de/fileadmin/lehrstuehle/ki/research/RapidMinerLODExtension/RapidMinerLODExtensionManual.pdf (Date: 15.12.2013).

Paulheim, H., Fürnkranz, J., 2012. Unsupervised Generation of Data Mining Features from Linked Open Data. In: International Conference on Web Intelligence, Mining, and Semantics (WIM's).

H. Rekow Einführung in das Data Mining. GRIN Verla–Publisher 2013.

Reichenberger, K., 2012. Kompendium semantischer Netz: Konzepte, Technologie, Modellierung. X.media.press. Springer.

T.A. Runkler Data Mining. Vieweg-Teubner Verlag, Wiesbaden. Saake, G., Heuer, A.,1999. Datenbanken–Implementierungstechniken, ISBN 2010 3-8266-0513-6.

Saravanan, D., Srinivasan, S., 2013. Matrix based indexing technique for video data. J. Comput. Sci. 9, 534–542.

Schmid, T., 2013. Wie man zwischen den Zahlen liest. Data-Mining und computergestützte Vorhersagen am Beispiel Bioinformatik. In: Arbeitstitel–Forum für Leipziger Promovierende 5, 13–29.

S. Shekhar P. Zhang Y. Huang Spatial Data Mining Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, Springer Science+Business M LLC 2010 837–854.

Tanner, R., 2013. Data Mining–das etwas andere Eldorado. Technologie IT-Methoden 8, 37–42.

Thushari, S., Vilas, W., Hitesh, N.S., 2012. Disaster mitigation and preparedness using linked open data. J. Ambient Intelligence Humanized Comput doi:http://dx.doi.org/10.1007/s12652-012-0128-9.

W.G. Touw, J.R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, S.A.F.T. Hijum. Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform 14, 315–326 2013 10.1093/bib/bbs034.

URL, 2014; http://www.adb.org/themes/urban-development.

Villa, F., Athanasiadis, I.N., Rizzoli, A.E., Emilio, A., 2009. Modelling with knowledge: A review of emerging semantic approaches to environmental modelling. Environ. Model. Software 24, 577–587.

Vintrou, E., 2013. Data Mining, A Promising Tool for Large-Area Cropland Mapping. IEEE J. Sel. Top. Appl. 99. 1–7.

Williams, G., 2011. Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery (Use R!), Springer Science Business Media, LLC,2011.

I.H. Witten ,F. Eibe. 2001. Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen. Carl Hanser Verlag München Wien.

I.H. Witten F. Eibe M.A. Hall Data Mining Practical Maschine Learning Tools and Techniques 2011 Morgan Kaufmann Publishers.

Wu, F., Zhan, J., Yan H., Shi C., Huang, J., 2013. Land cover mapping based on multisource spatial data mining approach for climate simulation: a case study in the farming-pastoral ecotone of North China, Advances in Meteorology, http://dx.doi.org/10.1155/2013/520803, (Date: 01.012.2013).

Yuan, J., Yang, M.,Wu, Y., 2011. Mining discriminative co-occurrence patterns for visual recognition. CVPR. 2777–2784.