International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 10 (2018) pp. 7835-7841

Wartortle - A Data Mining and Knowledge Discovery Suite for Data Analysis and Reporting

M.A. Shanti[a] and K.Saravanan[b]

[a]*Research Scholar, Department of Computer Science and Engineering, PRIST University, Thanjavur, India.*

[b]*Dean, Faculty of Computer Science, PRIST University, Thanjavur, India.*

Abstract

Data is generated from various sources ranging from microsensors to large-scale user access databases. These data are stored both physically and virtually. The analysis part of these huge volumes of data poses a very big challenge as the conventional technologies will not be able to process it. Data mining is the practice of examining large antecedent databases in order to generate new information. Knowledge Discovery is inter-disciplinary in nature. Knowledge Discovery is the process of obtaining meaningful information from a huge volume and variety of data. When such large data store has to be analyzed for business intelligence then state of the art technologies must be considered. In this paper, a general framework has been proposed which characterizes the Knowledge Discovery and Data mining. Then a platform named WARTORTLE (Work Analytics Report with Tasks Optimized for Requests Transmission enhanced with Link Encryption) has given which is based on the proposed framework. The research work also includes the design and its characteristics expressing the operational efficiency and efficacy of the framework proposed.

Keywords: Data mining and knowledge discovery, Data mining systems, Framework, Business Intelligence, Framework approach, procedural workflow.

INTRODUCTION

In recent years, two communities have grown around a joint interest on how big data can be exploited to benefit education and the science of learning: Educational Data Mining and Learning Analytics[1]. The advances in location-acquisition and mobile computing techniques have generated massive spatial trajectory data, which represent the mobility of a diversity of moving objects, such as people, vehicles, and animals. Many techniques have been proposed for processing, managing, and mining trajectory data in the past decade, fostering a broad range of applications[2]. The aim of any data mining technique is to build an efficient predictive or descriptive model of a large amount of data. Applications of evolutionary algorithms have been found to be particularly useful for automatic processing of large quantities of raw noisy data for optimal parameter setting and to discover significant and meaningful information. Many real-life data mining problems involve multiple conflicting measures of performance, or objectives, which need to be optimized simultaneously[3]. It sounds like mission impossible to connect everything on the Earth together via Internet, but the Internet of Things (IoT) will dramatically change our lives in the foreseeable future, by making many "impossibles" possible. To many, the massive data generated or captured by the Internet of Things are considered having highly useful and valuable information. Data mining will no doubt play a critical role in making this kind of system smart enough to provide more convenient services and environments[4]. Biomedical research is drowning in data, yet starving for knowledge. Current challenges in biomedical research and clinical practice include information overload – the need to combine vast amounts of structured, semi-structured, weakly structured data and vast amounts of unstructured information – and the need to optimize workflows, processes, and guidelines, to increase capacity while reducing costs and improving efficiencies. There is the pressing need to combine vast amounts of diverse data, including structured, semi-structured and weakly structured data and unstructured information. Interestingly, many powerful computational tools advancing in recent years have been developed by separate communities following different philosophies: Data mining and machine learning researchers tend to believe in the power of their statistical methods to identify relevant patterns – mostly automatic, without human intervention. There is, however, the danger of modeling artifacts when end-user comprehension and control are diminished. Additionally, mobile, ubiquitous computing and automatic medical sensors everywhere, together with low-cost storage, will even accelerate this avalanche of data[5]. Though the main efforts in the Knowledge Discovery and Data mining populace have been so steadfast to the development of efficient mining algorithms and techniques, some standardization initiatives, aimed at solving the Knowledge Discovery and Data mining problem as a whole, have been proposed. Here, a Knowledge Discovery and Data mining procedure are taken as a ”multi-step process where the input of one mining operation can be the output of another” and a data mining and knowledge discovery algebra is proposed. Thus, in the perspective of a unified approach, a data mining system should provide support to the entire Knowledge Discovery and Data mining life cycle within a unique, multi-step integrated environment, providing suitable and salient features to enable the construction of a Knowledge Discovery and Data mining analysis as an interactive and an incremental process. To this

end, there is a number of requirements that an integrated Knowledge Discovery and Data mining environment should meet, including:

(a) ability to integrate various data sources

(b) ability to employ a vast number of procedures and techniques

(c) denoting the affluence of data mining and knowledge discovery procedures and techniques

(d) abundant procedure for mining varied data sources

(e) supporting multivariate environments

(f) managing criteria with the support of an extension to various dimensions

(g) ability to create, monitor and improvise new techniques

Additional requirements that a Knowledge Discovery and Data mining system should fulfill in order to cope with real-world applications include:

i. Easy Graphical User Interface

ii. Support of Multiple Data Types

iii. Integration with Cloud

iv. Integration with Big Data environment

v. Enhancements pertaining to sensitive data handling like encryption

In the recent years, there are a lot of tools existing for the purpose of Knowledge Discovery and Data mining. These tools are available as commercially marketed and open source licensed. All these tools have a variety of options enabling a researcher to try out a lot of features on the datasets. Many commercial Knowledge Discovery and Data mining tools available in the market result from the integration of data mining techniques and procedures into commercial statistical tools, e.g., SPSS[6][9] and SAS[7], while others, such as Oracle Data Mining[8], result from the integration of data mining and Knowledge Discovery solutions into business intelligence products. tools like IBM SPSS Modeler[10][11][12] discover patterns and trends in structured or unstructured data more easily, using a unique visual interface supported by advanced analytics. It produces deeper insight and more accurate predictions by utilizing all of your data assets to create a complete view of your customers or constituents. IBM SPSS Modeler integrates with IBM Cognos 8 Business Intelligence software [13][14] and gives a complete range of advanced analytical functions, including state-of-the- art algorithms, automated data preparation and rich, interactive visualization capabilities. There are many surveys on Knowledge Discovery and Data mining tools[15][16][17][18]. Each survey work covers a specific line of research and gives the nature and features of the selected tools.

In this research work, a platform named WARTORTLE was proposed. It is a Knowledge Discovery and Data mining platform assisting the designer during the entire DMKD processes.WARTORTLE is visually interactive and it enables the structuring of the DMKD process as a technical and procedural workflow representing a balance of elementary operators on three worlds: the data, the dimension and the design worlds (the DDIDE Model). In the data world, the data types along with their attributes are taken into account. In the dimension world, the flow of data into the analytical pathways are considered. In the design world, the workflow of the analytics is given more importance. The main features of WARTORTLE are listed below.

a) Event listing

b) Scalability to handle high data traffic

c) Ease of customization and integration

d) Control of data at any given point of Analysis

e) Enabling a wide data range comparison

f) Analytics based on Geographical data can analyze with geographical metrics

g) Data conversion

h) Navigation Summary

i) Dashboard Integration

j) Plugin supported

k) Phase control (from target data selection to design deployment)

Currently, WARTORTLE was planned to work with relational databases. To give a concise view, the contribution of this research works are

I. A minimal algebraic model namely DDIDE Model has been proposed for data mining and knowledge discovery queries. The DDIDE Model is inspired from a lot of frameworks and algebraic models proposed over the years. The DDIDE model is based on the concept of integrating various data analysis factors in the DMKD process and give the researchers a basic foundation of problem-solving methodology.

II. The architecture which implements the DDIDE Model has been proposed

This research work is organized as follows. Section 2 of this paper defines the DDIDE Model. Section 3 develops and presents an overview of WARTORTLE platform. Section 4 gives the evaluation of the proposed analytical suite. Section 5 encapsulates the entire research work.

DDIDE MODEL – STRUCTURING THE KNOWLEDGE DISCOVERY AND DATA MINING PROCESS AS ALGEBRAIC PROCESSES

Mathematical modeling

Mathematical models are useful for a variety of reasons. Foremost, models represent the mathematical core of a situation without extraneous information[19]. Weak factorization systems are familiar in essence if not in name to algebraic topologists. Loosely, they consist of left and right classes of maps in a fixed category that satisfies a dual lifting property and are such that every arrow of the category can be factored as a left map followed by a right one. Neither these factorization nor the lifts are unique; hence, the adjective "weak." Two weak factorization systems are present in Quillen's definition of a model structure on a category[20].

Algebraic Modeling

Algebra is a generalization of arithmetic in which letters representing numbers are combined according to the rules of arithmetic. It can also be defined as any of various systems or branches of mathematics or logic concerned with the properties and relationships of abstract entities (as complex numbers, matrices, sets, vectors, groups, rings, or fields) manipulated in symbolic form under operations often analogous to those of arithmetic. An algebraic model takes a real-world situation described in words and describes that situation using algebra. An algebraic model uses variables and numbers. Algebraic models are usually easy to explore because sequence of values for the independent variable can be simply generated and plot the resulting values of the model's dependent variable.

DDIDE Model

In DDIDE model, there are three worlds viz Data, Dimension and Design. Each world, as indicated in the proposed model as phase, has its own criteria and attribute determinations.

Data World (DA-WORLD)

The Data world (DA-WORLD) consists of datasets with all their related and dependent attributes. These attributes can be given a factor and their dependencies can be correlated with these factors. Let's assume the DA-WORLD consists of four different datasets. The following table gives a clear picture of how the datasets are represented in the DA-WORLD.

Table 1: Representation of Data in the Data Phase of the DDIDE Model

Datasets	Nature of Data source	Data representation	Dataset representation	Data with their Database representation
A	infinite	A_{num}	$A=\{A_1,A_2,A_3,...\}$	$tDB(A_{num})$
B	infinite	B_{num}	$B=\{B_1,B_2,B_3,...\}$	$tDB(B_{num})$
C	infinite	C_{num}	$C=\{C_1,C_2,C_3,...\}$	$tDB(C_{num})$
D	finite	D_{num}	$D=\{D_1,D_2,D_3,D_4,D_5\}$	$tDB(C_{num})$

Dimension World (DIM-WORLD)

The Dimension world (DIM-WORLD) consists of the factors correlating to the datasets employed for analysis. The dimensions can be two or three depending on the factors relating to the analyzed dataset. Let's assume that the same four datasets used in the DA-WORLD are factored into the DIM-WORLD. Then the dimensions are given as follows.

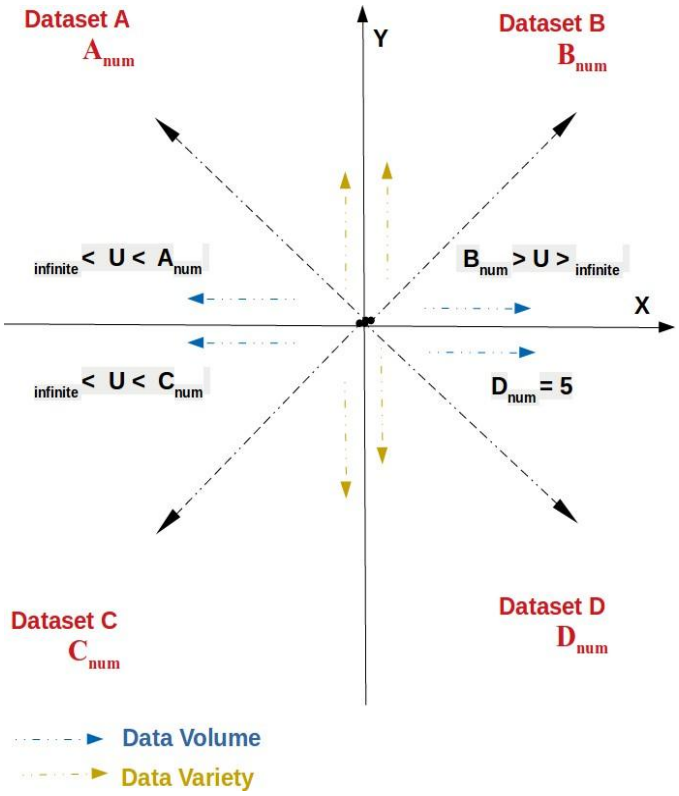

Figure 1: Dataset representation in the DIM-WORLD

Design World (DES-WORLD)

The Design world (DES-WORLD) consists of the patterns related to the data and their attribute values correlating to their respective analytical procedures. In the DES-WORLD, the databases with the datasets can be finite or infinite, are comprehended and given a model form in which the datasets get a complete representation of the factors, thus enabling the researchers to get a clear conceptual knowledge of how the data plays a definitive role in determining the business intelligence. Let's assume that the same four datasets used in the DA-WORLD which are given the dimensions in DIM-WORLD are factored into the DES-WORLD. Then the designs are given as follows

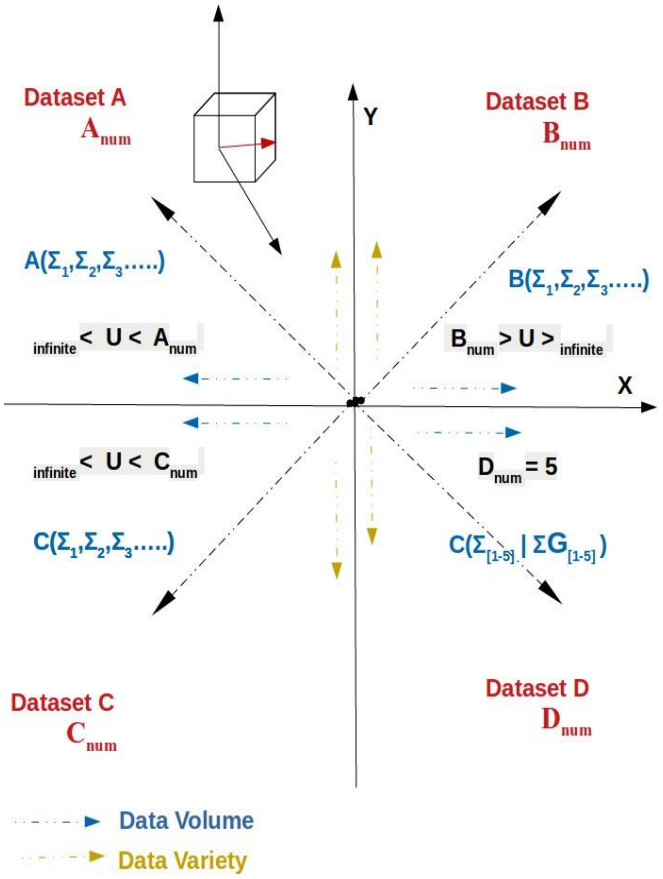

Figure 2: Dataset representation in the DES-WORLD

DDIDE model and its Minimal Algebraic computation

Let $Q \subseteq RE^{num}$

Where Q is the Database and it is given by the elemental attributes RE^{num} Let the given property of a dataset A be

$A=(A1,A2,.....A_X) \varepsilon RE^{num}$

determine if $A \varepsilon Q$, Then set

$Q = \{ (A1,A2,.....A_X) \mid \Pi (A1 - A_X) \neq 0 \} \subseteq RE^{num}$

Here Q value can be computed with the data attributes but the relationship between the data and its source can be computed only by determinations considering the fact that all datasets subjected to analysis are from the same databases. Then the new computation for Q will be given as

$Q = \{ (A1,A2,.....A_X)^{(DBD)} \mid \Omega (A1 - A_X) \neq 0 \mid \Pi (A1 - A_X)^{(DBD)} \neq 0 \} \subseteq RE^{num}$

$Q = \{ (A1,A2,.....A_X)^{(DBD)} \mid \Delta (A1 - A_X) \neq 0 \} \subseteq RE^{num}$

Where Ω is the nodal value, Δ is the sub tenor value calculated by adding the dataset attributes with the deployed attribute value which is a fixed value given by the analyst at the start of the data analysis process. The following figure depicts the DDIDE Model in detail with all its encompassing processes.

$$Q(A) = \sum_{A=0}^{\infty} \frac{f^{(i)}(0)}{i!} A^i \gg f(A) = \frac{1}{\sigma\sqrt{2\pi}} DB^{-\frac{(A-\mu)^2}{2\sigma^2}}$$

Where Ω is the nodal value, Δ is the sub tenor value calculated by adding the dataset attributes with the deployed attribute value which is a fixed value given by the analyst at the start of the data analysis process. The following figure depicts the DDIDE Model in detail with all its encompassing processes.

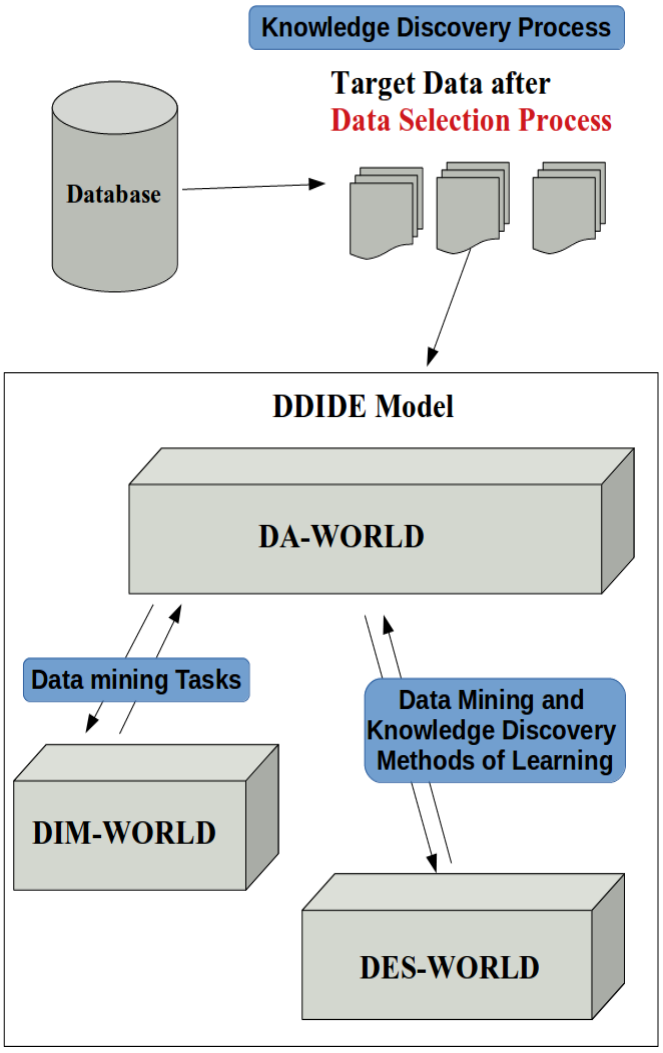

Figure 3: DDIDE Model

Comparison of DDIDE model with existing 3W Model

The 3W Model[21] emphasizes the underlying philosophy of integrating many factors of knowledge to solve the problem statement. The following table gives a comparison between the 3W Model and our proposed DDIDE model.

Table 2: Comparison between 3W Model and DDIDE Model

Factors	3W Model	DDIDE Model
Data Type support	Variable	Variable
Data Mining Analysis support	Minimal. Few exceptions due to clustering algorithms	Complete
Knowledge Discovery	Minimal. Few exceptions due to	Complete
support	clustering algorithms	
Operators support	Built in support	Customization based upon requirements
Computation migration support	Minimal	Any
Data migration support	Minimal	Any
Computation migration support	Minimal	Any
Model operators	Algebraic	Comprehensive as it is open support framework model
Mining operators	Region-based	Comprehensive as it is open support framework model
Models	Tables, mostly multidisciplinary relational	Comprehensive as it is open support framework model

AN OVERVIEW OF WARTORTLE

The proposed framework WARTORTLE is a completely graphical environment for designing and managing. Data Mining and Knowledge Discovery processes as algebraic expressions based on the DDIDE Model. The following figure 4 describes the Architecture of WARTORTLE.

The work analysis of each Architectural component are listed below:

(a) **Platform Support:** The basic functionality of the Platform support is to enhance the data reading ability by enabling data input from varied sources viz Physical drives, cloud storage, files stored in removable media., etc,.The the main function of the Platform support is to provide data accountability at any given point of time to the Reader module.

(b) **Reader:** The main function of this section is to take data from the platform and enable the readability criteria for Storage and Dashboard. Reader finalizes the data processing procedure by altering the flow of data control and inhibiting any false positives while data transfer.

(c) **Storage:** The main function of this section to make the data available for retrieval at any given point of time for the other sections. The Storage section has separate modes for various data stores and high availability of data is made as for its primary concern. The data store differs according to the data source and the data source availability and integration with the platform support determine the efficiency of the Storage. The Storage section supports query based retrieval for better support.

(d) **Run Time Engine:** Runtime engine lays the software framework to build and create data analysis processes. They provide features from data assimilation to artificial intelligence. Runtime engine is responsible for rendering graphics(Dashboard), collision detection(Network), memory management(Storage and Processing), and much more options in the architecture.

(e) **Data:** Data section partitions these massive data and executes queries efficiently on these that involve groupings, look-ups, orderings, and complex selections.

(f) **Design:** Design section models the data with its respective attributes and is responsible for the data segregation and aggregation process. When a data is subjected to analysis in the WARTORTLE, the design section determines the extent to which the data can be analyzed and it is based on the design section decision, the dimension section segregates the attributes of the data and store it in the storage for later retrieval.

(g) **Dimension:** The dimension section differentiates the attributes of all the data and makes it in a query readable format for the design section. This query readable section is partially given by the design section as it determines the data characteristics.

(h) **Workload Management:** In this section, the entire data analysis part is taken into account and the process can be optimized at any given point of analysis time. The entire workflow is divided into seven perspectives and is seeded to the dashboard section for better researcher decisions. The seven perspectives of the Workload management are as follows:

i. Source perspective
ii. Storage perspective
iii. Design perspective
iv. Engine perspective
v. Load perspective
vi. Workflow perspective
vii. Performance perspective

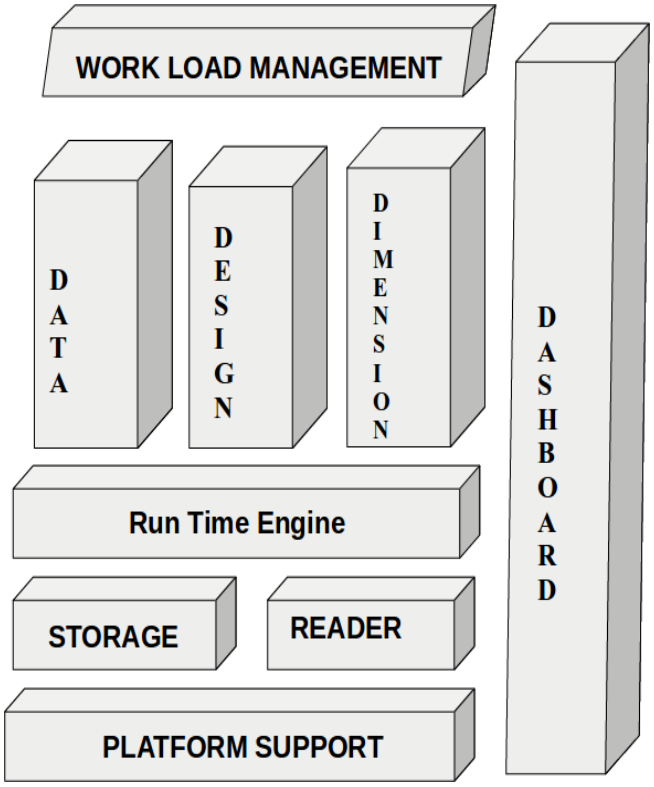

Figure 4: WARTORTLE Architecture

Again, the open architecture model enables the conceptualization of the ad hoc tasks which has the capability to handle and present user-specific tasks. Technically, the tasks that can be done in WARTORTLE is of ten types namely:

(1) Analytical tasks

(2) Storage tasks

(3) Sorting tasks

(4) Variable Employ-ability task

(5) Prediction task

(6) Classification task

(7) Structural task

(8) Rule Mining task

(9) Pattern task

(10) Time series task

EXPERIMENTAL ANALYSIS OF WARTORTLE

In this section, the evaluation of the WARTORTLE architecture was reported and a model of WARTORTLE as a web application was developed and assessed its performance. The following table gives a clear picture of WARTORTLE evaluation report.

Table 3: Evaluation of WARTORTLE

Criteria	Description
Data Management	Able to handle large datasets without any initial changes
Data Source management	WARTORTLE acquires datasets in Excel and CSV formats.
Data mining algorithm support	WARTORTLE architectural framework can be made to support most of the data mining algorithms and knowledge discovery procedures
Data visualization	WARTORTLE supports some graphical tools. In future works, an inbuilt graphics generation has been planned.
Open Community	WARTORTLE is not an open source project. But in future, there are plans to open coding support for some algorithms.

CONCLUSION

In this paper, first a general framework for modeling Data Mining and Knowledge Discovery processes has been proposed, called DDIDE Model, a variant of the 3W Model. According to this model, the essence of a Data Mining and Knowledge Discovery processes can be regarded as the interaction between three various independent worlds: the data, dimension and the design world. This way, any Data Mining and Knowledge Discovery process can be formalized as an algebraic expression, that is essentially a composition of operators representing 1146 (pseudo)elementary operations on the two worlds. Then, Data Mining and Knowledge Discovery platform has been described called WARTORTLE. It provides a visual dashboard and gives provides support to all steps of all DMKD process, from target data acquisition to design and dimension exploration with analytical deployment. In future, the WARTORTLE can be enhanced to handle all algorithms and will also support plugins and application programming interface support to make it interoperable with other applications.

REFERENCES

[1] Baker, Ryan Shaun, and Paul Salvador Inventado. "Educational data mining and learning analytics." In *Learning analytics*, pp. 61-75. Springer New York, 2014.

[2] Zheng, Yu. "Trajectory data mining: an overview."*ACM Transactions on Intelligent Systems and Technology (TIST)* 6, no. 3 (2015): 29.

[3] Mukhopadhyay, Anirban, Ujjwal Maulik, Sanghamitra Bandyopadhyay, and Carlos Artemio Coello Coello. "A survey of multiobjective evolutionary algorithms for data mining: Part I."*IEEE Transactions on Evolutionary Computation* 18, no. 1 (2014): 4-19.

[4] Tsai, Chun-Wei, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang. "Data mining for Internet of

Things: A survey."*IEEE Communications Surveys and Tutorials* 16, no. 1 (2014): 77-97.

[5] Holzinger, Andreas, and Igor Jurisica. "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions." In *Interactive knowledge discovery and data mining in biomedical informatics*, pp. 1-18. Springer Berlin Heidelberg, 2014.

[6] Green, Samuel B., and Neil J. Salkind. *Using SPSS for Windows and Macintosh: Analyzing and understanding data*. Prentice Hall Press, 2010.

[7] SAS Institute. *SAS/STAT 9.1 User's Guide the Reg Procedure:(Book Excerpt)*. SAS Institute, 2008.

[8] Tamayo, Pablo, Charles Berger, M. Campos, Joseph Yarmus, B. Milenova, Ari Mozes, Margaret Taft et al. "Oracle data mining."*Data mining and knowledge discovery handbook* (2005): 1315-1329.

[9] Nie, Norman H., Dale H. Bent, and C. Hadlai Hull. *SPSS: Statistical package for the social sciences*. No. HA29 S6. New York: McGraw-Hill, 1970.

[10] Modeler, IBM SPSS. "Algorithms Guide."*IBM Corporation* (2012).

[11] Modeler, IBM SPSS, and Algorithms Guide. "IBM Corporation."*NY, USA* (2010).

[12] McCormick, Keith, Dean Abbott, Meta S. Brown, Tom Khabaza, and Scott R. Mutchler. *IBM SPSS modeler cookbook*. Packt Publishing, 2013.

[13] Volitich, Dan. *IBM Cognos 8 Business Intelligence*. mitp, 2009.

[14] Grabova, Oksana, Jerome Darmont, Jean-Hugues Chauchat, and Iryna Zolotaryova. "Business intelligence for sGoebel, Michael, and Le Gruenwald. "A survey of data mining and knowledge discovery software tools."*ACM SIGKDD explorations newsletter* 1, no. 1 (1999): 20-33. mall and middle-sized entreprises."*ACM SIGMOD Record* 39, no. 2 (2010): 39-50.

[15] Goebel, Michael, and Le Gruenwald. "A survey of data mining and knowledge discovery software tools."*ACM SIGKDD explorations newsletter* 1, no. 1 (1999): 20-33.

[16] Romero, Cristobal, and Sebastian Ventura. "Educational data mining: A survey from 1995 to 2005."*Expert systems with applications* 33, no. 1 (2007): 135-146.

[17] Kurgan, Lukasz A., and Petr Musilek. "A survey of Knowledge Discovery and Data Mining process models."*The Knowledge Engineering Review* 21, no. 1 (2006): 1-24.

[18] Mitra, Sushmita, Sankar K. Pal, and Pabitra Mitra. "Data mining in soft computing framework: a survey."*IEEE transactions on neural networks* 13, no. 1 (2002): 3-14.

[19] Mangasarian, Olvi L. "Mathematical programming in data mining."*Data mining and knowledge discovery* 1, no. 2 (1997): 183-201.

[20] Quillen, Daniel. "Elementary proofs of some results of cobordism theory using Steenrod operations."*Advances in Mathematics* 7, no. 1 (1971): 29-56.

[21] Johnson, Theodore, Laks VS Lakshmanan, and Raymond T. Ng. "The 3w model and algebra for unified data mining." In *VLDB*, pp. 21-32. 2000.