BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **19**, No 3
Sofia • 2019 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2019-0023

Towards Big Data Analytics in the e-Learning Space

Ivan P. Popchev[1], *Daniela A. Orozova*[1,2]

[1]*Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria*
[2]*Faculty of Computer Science and Engineering, Burgas Free University, 62 San Stefano Str., 8001 Burgas, Bulgaria*
E-mails: ipopchev@iit.bas.bg didyorozova@gmail.com

***Abstract*:** *The issues related to the analysis and management of Big Data, aspects of the security, stability and quality of the data, represent a new research, and engineering challenge. In the present paper, techniques for Big Data storage, search, analysis and management in the area of the virtual e-Learning space and the problems in front of them are considered. A numerical example for explorative analysis of data about the students from Burgas Free University is applied, using instrument for Data Mining of Orange. The analysis is a base for a system for localization of students at risk.*

***Keywords*:** *Big Data Analytics, Data Mining, Map/Reduce, e-Learning, Orange.*

1. Introduction

The global and all-penetrating character of the Internet of things and the dynamic interaction of the connected objects assume exponential increase of the volume and the accessibility of data that is characterized by complexity, heterogeneity, high incoming speed and lack of structure, introducing the Big Data paradigm.

Big Data is a novel and developing concept which describes huge in volume data with different degree of structuring which can be analyzed with the help of highly productive methods for finding of trends, models and associations in the data [3]. A group of technologies and methods for processing of Big Data is connected to the concept of Big Data Analytics. The problems that the IT specialists in the field are faced with include gathering, storage, search, sharing, transfer, analysis and visualization of the data. The characteristics of the Big Data are described with the help of five V's [7]:

- Volume – large volume of data that are in need of processing.
- Velocity – high speed of generation and analysis of flows of data.
- Variety – the data come in different shapes and from different sources.
- Veracity – the quality and the precision of the data varies.
- Value – usability and usefulness of the data.

One of the challenges in front of Big Data and their analysis is related to the structuring and the storage of heterogeneous data, which come in with high velocity. The sources of data are usually of different type, the generated data are subject to various standards. This imposes additional steps for transformation of the data in a form, which satisfies the defined rules in the platform. According to the degree of structuring, the data are divided into three classes: structured, non-structured and semi-structured data. An example of structured data are tables, databases, reports, etc. Non-structured data are generated by: Internet of things, internet of people (social networks – Facebook, Tweeter, LinkedIN, etc.); Internet of Things; Internet of Location (mobile phones, smart phones, tablets, etc.). The term semi-structured data is related to data stored in format XML, JSON or others. The different types of data require different approach and software for processing [4]. The variety of the forms, the way of creation and the type of the digital documents make them hard for structuring, indexing and search.

The technologies for storage of data and knowledge are diverse. Generally, the Data repository represent a space for storage of information with minimal functionality for analysis and search. They use different technologies for storage and access to the data such as the technology *Key-Value* (the database is a totality of ordered pairs of the type ⟨Key, Value⟩, *Document stores* (used in the work with documents in structured and non-structured format), *Graph stores* (store and analyze graph structures while making records of the semantics of the data, connections and relations between them). The modern applications have to store and process a set of other data – drawings, maps, multimedia, WWW, etc. They require processing of different structures. These requirements are satisfied up to a certain degree by the object-oriented, object-relational and other types of data management systems.

2. Search for information and knowledge

2.1. Internet technologies and algorithms for searching

The algorithms for searching of information and knowledge support users of Internet to find certain contents using keywords with respect to shape and type in a specific context in a concrete information environment. The development of the means for searching continues with the more and more wide penetration of complex information object (video, picture, sound, and multimedia). In practice, the following strategies for searching are used [1]:

- *meta*-searching – the approach is based on meta categories which allow the users to determine the focus of searching;
- *hierarchical* searching – the data is organized in a certain hierarchy which allows for in-depth searching of the storage;
- *content* searching – the user enters terms, keywords or text strings and in return receives a result with the found matches and evaluation of the relevance;
- *attribute* searching – requires the introduction of values which are compared with similar values from documents or other data. Having in mind such data retrieving, its organization and input of specific labels or attributes, connected to the

stored data, are of special importance. The attributes can be of various types: attributes for activity (connection with organizational activities); attributes for a field (tags with data according to the subject area); attributes for form (determine the physical representation of objects or the form); attributes for type (for example, procedure, manual, protocol, report, analysis, notes, good practice, solved problem, etc.); attributes for product (specifies the product or the service which the object refers to); attributes for time; attributes for place (showing the location), etc.

2.2. Big Data Analytics and Data Mining

Nowadays, with the invasion of the Internet of Things in our life the bind of the modern technologies with Big Data is of special significance. The need for data analysis for extraction of useful information increases. The terms Big Data, Big Data Analytics и Data Mining describe the Big Data as well as the technologies for collecting, processing, management of data and methods for analysis

Data Mining is the process of searching of hidden data and objective laws, initially unknown, non-trivial and practically useful, which are required for decision-making in different spheres of human activities. Here the emphasis falls not only on the extraction of new facts but also on the generation of hypothesis that can be checked.

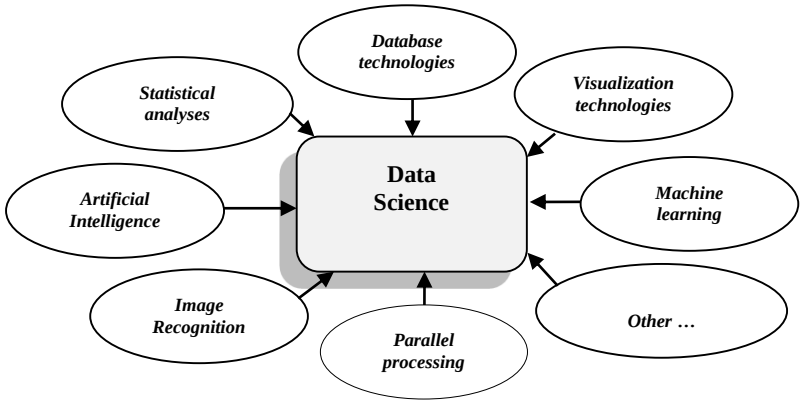

Fig. 1. Functional connections of Data Science

Big Data Analytics represents a development of the concept of Data Mining. It is also a development of the problems solved, application areas, data sources, methods of technological processing. It can be said that Data Mining is “alloy” of different disciplines and technologies for extraction of knowledge and data [2]. When, however, the scheme is complemented with technologies resulting from the requirements of the 5 V’s, it reflects the functional links of *Data Science* (Fig. 1).

Data science [9] allows for the combination of various approaches including techniques connected to data analysis in the field of statistics, machine learning, artificial intelligence, programming, communications, etc. The science of data includes the processes of data clearing and integration, selection and transformation of data, knowledge extraction, its analysis, evaluation and representation.

In the period from the appearance of the concept of Data Mining to the advance of Big Data, the volume of the analyzed data changes, high performance systems appear, new technologies, including Map/Reduce and its multiple program implementations.

Map/Reduce is a basic paradigm for distributed computations of huge arrays of data. Map/Reduce is used for representation of computing problems and simplifies the programming process, guarantees scalability, stability to failures and allows effective loading and splitting of data with the aim of its parallel service [4].

The work of the Map/Reduce algorithm consists of two steps: *Map* and *Reduce*. The basic idea is the input files to be separated into *M* parts. The separate input parts are distributed automatically by the master program between one or more machines or other parts of the program (workers). They receive input data and tasks for preliminary processing. The user assigns the *Map* function, which processes the input data and generates compositions of intermediate pairs of type ⟨key : value⟩. The results of the *Map*-step are recorded in intermediate files. The resulting intermediate lists with keys are grouped in R parts and are processed independently through the *Reduce* procedures. In this way, through the *Reduce* step the data from the preliminary processing shrinks. The main node receives the answers from the working nodes and generates the result based on them [14]. A huge variety of program implementations of the Map/Reduce technology exists. The concrete choice depends on the environment and on a collection of other factors.

3. Data Mining tools in the education space

During the recent years, we have witnessed a large use of Data Mining techniques for analysis and prognosis in various fields of contemporary life. Their aim is finding patterns, correlations and tendencies in huge volumes of data with the help of statistical and mathematical means and methods and methods of the artificial intelligence. Our efforts are aimed at integration of the Data Mining techniques with systems for electronic learning [5, 12]. Based on the accumulated data from the work of the system for electronic learning with different users, applying tools from the area of knowledge extraction from the data, different decisions can be made such as:

- to construct optimal learning environments with possibilities for personalized acquiring of key knowledge, abilities and competencies;
- to optimize the techniques for selection of test elements and suitable type of intervention in the activities of the student;
- to search for tendencies regarding the development of the processes of electronic learning and its service;
- to identify student types who are to be offered suitable continuation of their learning;
- to predict the students for whom there is a danger of failing to cope with the learning process;

- to perform analysis about the degree of acquiring and forgetting of knowledge for different time periods and for different types of problems as well as comparison of the indicators through the years.

The integration of the education spaces with Data Mining tools is necessary for the process of personalization of courses for electronic learning. On the basis of the obtained results, additional measures for analysis and change of learning courses can be introduced. This, on its part, is a path towards the increasing of the quality of learning.

Specialized personal assistants [13] can determine and measure a collection of personal qualities of the students which have an impact on the learning. Data from any given user can be analysed to establish and measure the user's personal skills that affect learning, like: curiosity (the user's desire to learn something new), persistence (even if a satisfactory solution exists, the user has the ability to peruse better solutions), connection (the ability of the user to make associations between data which are unrelated), complexity (the user's ability to process high quantity of information), as well as other personal skills.

A specific environment is developed in [10], which automatically generates tests, used for education of students of software engineering at the Faculty of mathematics and Informatics of the Plovdiv University and is a part of a virtual environment for education. The test generating is supported by special ontologies which are serviced by two intelligent agents named *Operative assistant* and *Estimating assistant.* The purpose of the first is to generate the test by forming random questions on a given topic using a knowledge base. The second, examines the answers of the users and keeps the records of the results during the test. It uses UML ontology for the examination of the answers. The two assistants are implemented as intelligent agents and each one of them has his own tasks. Both agents are developed in Java and JADE (Java Agent Framework), integrated in the Eclipse environment. They have their own sensors and effectors that are used for the accomplishment of interaction with their environment. The environment integrates different resources: Sharable Content Objects (SCO elements), e-Packets in SCORM 2004 (Sharable Content Object Reference Model), database with test questions and statistics for the students.

The model for generation of tests is developed in [11] and a specific approach to the representation of knowledge for education is proposed. The following three levels represent the model: *Domain level* (the basic building blocks are modelled as a container of interconnected education units); *Extractor level* (on this level, building blocks can be chosen according to the desired test structure); *Generator level* (on this level, the generation of test questions is modelled on the base of the extracted blocks). The knowledge base keeps structured education content in the form of ontology which represents the basic notions, the relations between them and the basic rules. The test questions and the results from the tests are kept in a relational database. The generated tests contain different types of questions in correspondence with the QTI standard. This standard is chosen because it offers well structured specification of e-learning materials such as the questions' structure, grades and results in XML format.

During the generation of questions, a set of suitable axioms is extracted from the ontology with the help of specific criteria which depend on the test generated.

A Generalized Net (GN) model of the processes related to the choice of electronic test system and its follow-up work is presented in [8]. In the paper, the model from [6] is modified to fit with the process of learning with intelligent system of different style of tutoring (e-Teacher, e-Trainer, learning game, etc.) and the models of automatic test generation from [10, 11] are applied. The focus is on the mechanism for selection of an appropriate type of system for knowledge and skills evaluation of the students.

4. An example for data analysis of the student's activities at the Burgas Free University

Different algorithms and techniques for classification are used for the data analysis of the students, during their education at the Burgas Free University. The software chosen for processing and analysis of the data is Orange [15], development of the Faculty of Bioinformatics at the University of Ljubljana. It is an open source software based on Python and offers a sequence of instruments for Data Mining and visualization of data. It works with data from different formats (CSV files, Excel sheets, SQL tables, and data from URL). For example, we select the file with data about a part of the students and load it with the help of the instrument *File*.

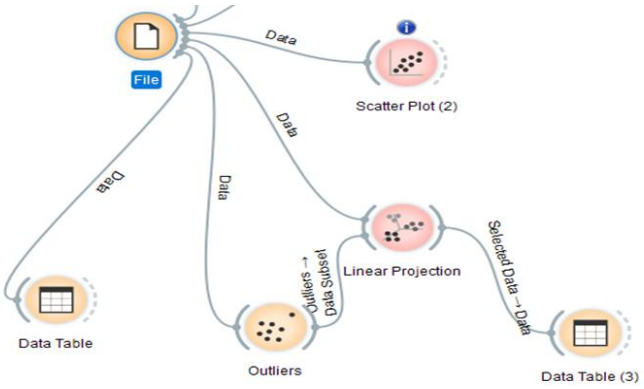

Fig. 2. Sample working process through the Orange system

As a result, information about the collection of data is visualized: size, number of records and types of data. Examination of data is made with the instrument *Data Table*, while its graphical representation is shown with the instrument *Scatter Plot* (Fig. 2). Here we could also filter the data with the help of the instrument *Select Rows* and to identify the columns on which the analysis will be made. In this case, initially we check if there are values which differ significantly from the others with the instrument *Outliers*.

The cluster analysis allows for identification of subsets of data with common characteristics. For the purpose of this analysis, the instrument *Distance* is included

for the measurement of the distance between separate points and we choose metrics for the distance with the instrument *Distance Metric* (in this case – *Euclidean* distance).

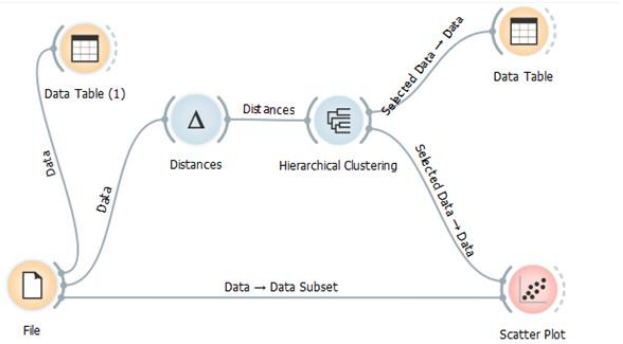

Fig. 3. Process of hierarchical clusterization and visualization of the results

Creation of the link *File* and *Distances* follows and hierarchical clusterization is applied with the help of the instrument *Hierarchical Clustering* (Fig. 3). Finally, a report can be made with the instrument *Report*.

First experiments were made by a small data sample limited by initially available real data for students (399 records). The instrument *Hierarchical Clustering* carries out a hierarchical grouping of arbitrary types of object through the evaluated distances *Distances* and shows the corresponding dendrogram.

Dendrogram is a graph tree in which every node represents one step of the process of forming of the clusters. It carries additional information about the distance between two clusters. In this case, a hierarchical cluster is built with depth 5 and in every cluster students are grouped according to the grade from the diploma. Students who have grade 0 as a value from the exam/matriculation are separated in the first subgroup. Most probably, these are students who have not taken exam or/and matriculation exam due to the changing laws during the years. The result from the cluster analysis is represented in Fig. 4. The data is visualized according to the faculty number of the students.

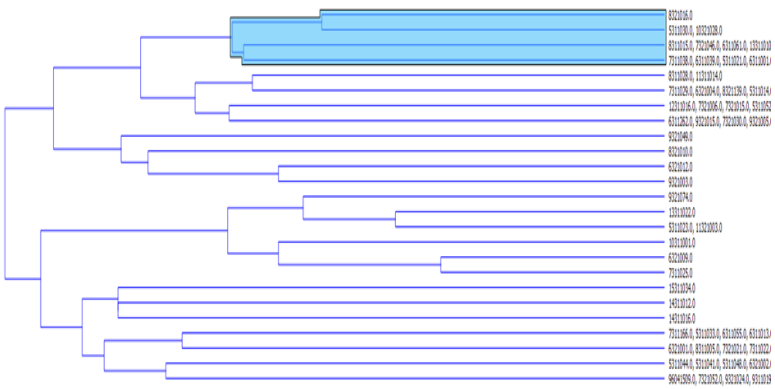

Fig. 4. Dendrogram – result of the work of the instrument Hierarchical Clustering

The instrument supports four ways of measurement of distances (*Linkage*) between the clusters (*Single linkage, Average linkage, Weighted, Complete linkage*). With the instruments of the system Orange, every cluster can be analyzed, described and visualized.

The graphical user interface of Orange gives opportunity to the users to concentrate on the scientific analysis of the data instead of on the coding of the algorithms. The system also offers components for machine learning and possibilities for extension of the functionality of the data extraction from outer sources, text extraction, network analysis, etc.

The science of data in favor of teaching is a new view which combines a collection of approaches to the processing of data obtained from various universities, environments and systems. Of particular interest is the data about the students, dropping out of the online courses or the distance form of education. Here, based on the collected data about the students in various forms of electronic education, we look for a connection between statistical methods, machine learning, discovery of behavioral models and data analysis. For the purpose of this research, an instrument identifying and predicting the reasons for falling behind or dropping out of the students is developed. A huge quantity of data is gathered on a daily basis, which we process through an open source distributed processing framework – Hadoop, based on Map/Redude Algorithm.

The instrument processes the data that is currently gathered during the students' education in their courses. Apart from that, we also use data from survey study. The survey is sent via e-mail in the third week of the course. In the base of the survey lie questions stimulating the students to determine the level of their coherence with the respective discipline as well as questions which determine their opinion about the level of difficulty of the subject. The data obtained from the surveys add a series of new characteristics which are directly connected to the dropping out of students such as lack of interest of the student, lack of time, organizational obstacles during the education, etc.

With help of the developed system for monitoring and collecting of data for the students and their activity (going over electronic text and video materials, participation in forum and/or groups based on interests, etc.) our efforts are aimed at searching for an approach that gives good results in the localization of students at risk. This is a multi-step procedure which is directly connected to the Big Data Analytics in the e-Learning Space.

5. Conclusion

The Big Data direction concentrates the efforts towards the organization, storage, processing and analysis of huge arrays of data. The processing and storage of Big Data requires a new view and joint application of several established technologies. Some open problems in this direction are related to the optimization of the access to Big Data using agents for knowledge extraction and new techniques for data analysis. Our future research will be towards the applications of Big Data in the field of the virtual learning spaces, as well as the search for models and associations on data,

collected during the dynamic interaction of the objects in the Internet of Things ecosystem.

References

1. Alexandrov, F., L. Egorova, L. Gokhberg, A. Myachin, G. Sagieva. Pattern Analysis in the Study of Science, Education and Innovative Activity in Russian Regions. – Computer Science, Vol. **17**, 2014, pp. 678-694.
2. Bernard, M. Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance. John Wiley & Sons Ltd., New York, 2015.
3. Beuerlein, B., at al. Big Data and the Role of the Actuary. American Academy of Actuaries, Washington, June 2018.
4. Fox, G. Big Data HPC Convergence and a Bunch of Other Things. **http://www.slideshare.net/Foxsden/big-data-hpc-convergence-and-a-bunch-of-other-things**
5. Glushkova, T., S. Stoyanov, I. Popchev, S. Cheresharov. Ambient-Oriented Modelling in a Virtual Educational Space. – Compt. Rend. Acad. bulg. Sci., Vol. **71**, 2018, No 3, pp. 398-406.
6. Orozova, D., K. Atanassov. Generalized Net Model of the Process of Selection and Usage of an Intelligent e-Learning System. – Compt. Rend. Acad. bulg. Sci., Vol. **65**, 2012, No 5, pp. 591-598.
7. Orozova, D., K. Atanassov. Generalized Net Model of Processes Related to Big Data. – Compt. Rend. Acad. bulg. Sci., Vol. **71**, 2018, Book No 12, pp. 1679-1686.
8. Orozova, D. Appropriate e-Test System Selection Model. – Compt. Rend. Acad. bulg. Sci., Vol. **72**, No 6, pp. 811-820.
9. Peng, R., E. Matsui. The Art of Data Science. Skybrude Consulting, LLC, 2016. **http://leanpub.com/artofdatascience**
10. Stancheva, N., A. Stoyanova-Doycheva, S. Stoyanov, I. Popchev, V. Ivanova. An Environment for Automatic Test Generation. – Cybernetics and Information Technologies, Vol. **17**, 2017, No 2, pp. 183-196.
11. Stancheva, N., A. Stoyanova-Doycheva, S. Stoyanov, I. Popchev, V. Ivanova. A Model for Generation of Test Questions. – Compt. Rend. Acad. bulg. Sci., Vol. **70**, 2017, No 5, pp. 619-630.
12. Stoyanov, S., I. Ganchev, I. Popchev, M. O'Droma. An Approach to the Development of Infostation-Based e-Learning Architectures. – Compt. Rend. Acad. bulg. Sci., Vol. **61**, 2008, No 9, pp. 1189-1196.
13. Stoyanov, S., V. Valkanov, I. Popchev, A. Stoyanova-Doycheva, E. Doychev. A Model of Context – Aware Agent Architecture. – Compt. Rend. Acad. bulg. Sci., Vol. **67**, 2014, No 4, pp. 487-496.
14. MapReduce parallel & distributed programming model, 11-06-2019, e-Book. **https://www.tutorialspoint.com/map_reduce/**
15. Orange. **https://orange.biolab.si/training/introduction-to-data-mining/**

Received: 21.06.2019; Second Version: 25.07.2019; Accepted: 05.08.2019