УДК 004.422

WEB-APPLICATION TO THE DATA MINING SYSTEM FOR EDUCATION AND RESEARCH

D.Y. Pertsau
Senior Lecturer of BSUIR

N.A. Iskra
Senior Lecturer of BSUIR

Belarussian State University of Informatics and Radioelectronics
E-mail: pertsev@bsuir.by.

Д.Ю. Перцев
Магистр технических наук, старший преподаватель на кафедре ЭВМ БГУИР, научный сотрудник НИЛ 3.6 НИЧ БГУИР.

Н.А. Искра
Магистр технических наук, старший преподаватель на кафедре ЭВМ БГУИР.

Abstract. In this paper the recently modernized system, that was previously developed and deployed on the computer cluster by ECM department, BSUIR, is described as a showcase of the intelligent data analysis by means of cloud computing. New results of research in image analysis with the help of this system are discussed.

Keywords. Intelligent data analysis, Cloud computing, Private cloud, Computer cluster, Image captioning.

Introduction

As a result of system modernization in 2019, BSUIR computing cluster is intended to be a private cloud built on SaaS model and consists of 7 Blade-servers, one of which is a control server and 6 others are computing servers, interconnected via InfiniBand4x QDR (40 Gbps) bus. Each computing module is equipped with two Intel Xeon E5-2650 processing units, 32 Gb RAM with DDR3 and two NVIDIA Tesla M2075 graphic processing units with 6 Gb RAM.

The near-term prospect is to provide services such as access to the data mining system for education and research developed by the ECM department and RL 3.6 research team [1-3] and Apache Zeppelin [4].

The main advantages of using BSUIR computing cluster as computing cloud are:

– self-sufficiency;

– access to local network of BSUIR, which allows to utilize the cluster by students and researchers, master and PhD students in particular, of any department; it is also considered to provide Internet access to the system in the nearest future;

– computational powers adequate to heavy and long-term processing load.

Services provided by BSUIR computing cluster.

Data mining system for education and research.

The system is a composition of layers:

– the libraries of algorithms services layer;

– the data analysis algorithms layer;

– Web-interface.

Services layer.

There are various libraries (frameworks) of ready-to-use data mining algorithms. The most popular among them are: scikit-learn [5], MLlib [6], Theano [7], Weka [8].

Each of the frameworks provides its own API and uses a number of programming languages. Moreover, different implementations can vary in efficiency for certain tasks. In case the developer needs algorithms from several libraries it becomes hard and requires high and diverse qualification.

The principal goal of the first layer is to unify the access to the algorithms.

The main purposes of the services layer are:

– to plug in a required library or to inform of the configuration errors;

– to check input parameters;

– to call a requested algorithm and to handle exceptions;

– to provide execution information along with error messages.

The data analysis algorithms layer.

The next layer is an interlay to provide access to the services. Its main responsibilities are:

– to store the information about connected libraries;

– to store the information about supported algorithms;

– to aggregate and arrange all the data;

– to communicate with the Web-interface.

Web-interface.

The final layer is the Web-interface, which is deployed on the ECM department's server. It's main tasks are:

– to provide the easy access to supported data mining algorithms;

– to build the chain of algorithms of necessary configuration for certain problems solution;

– to provide the results;

– to control access permissions.

Apache Zeppelin

Apache Zeppelin project is an open source Web-based notebook, which enables interactive data analysis.

The main adavantage of the project is that it supports the complex cloud computing infrastructure. Based on computing cluster of BSUIR and Apache Zeppelin, Web-interface access to such programming languages as Python and R is provided. Moreover, such technologies as Apache Spark [9] and TensorFlow [10], intended for data mining and deep learning correspondingly, are also supported.

Web-application for intelligent data analysis.

As a result of user-website interaction the JSON-description [11] of the operation pipeline with all the necessary parameters is formed and passed to the server. Server analyses the description and by means of REST-requests the data are further transferred to the services layer to be executed.

Web-interface provides the following operational pages:

– user profile;

– lists of projects for the certain user;

– project design;

– results.

After the authorization user is automatically redirected to previously created projects (figure 1). Functionally the page is divided into 2 parts:

– all enabled functionality is shown on the left;

– the content depending on the chosen action is shown on the right.

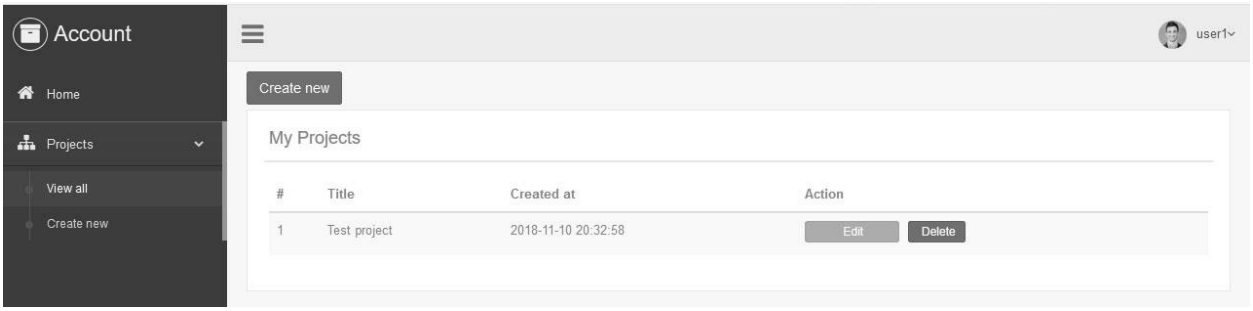


Figure 1. The list of projects

During the page loading REST-request for the list of the authorized user, which will be displayed dynamically on the screen, is created.

The project design is split into 3 steps:

– input data sampling;

– algorithm pipeline configuration;

– result display configuration.

During the input data sampling (figure 2) two data sources are available – Internet link to the CSV-file or CSV-file on the server.

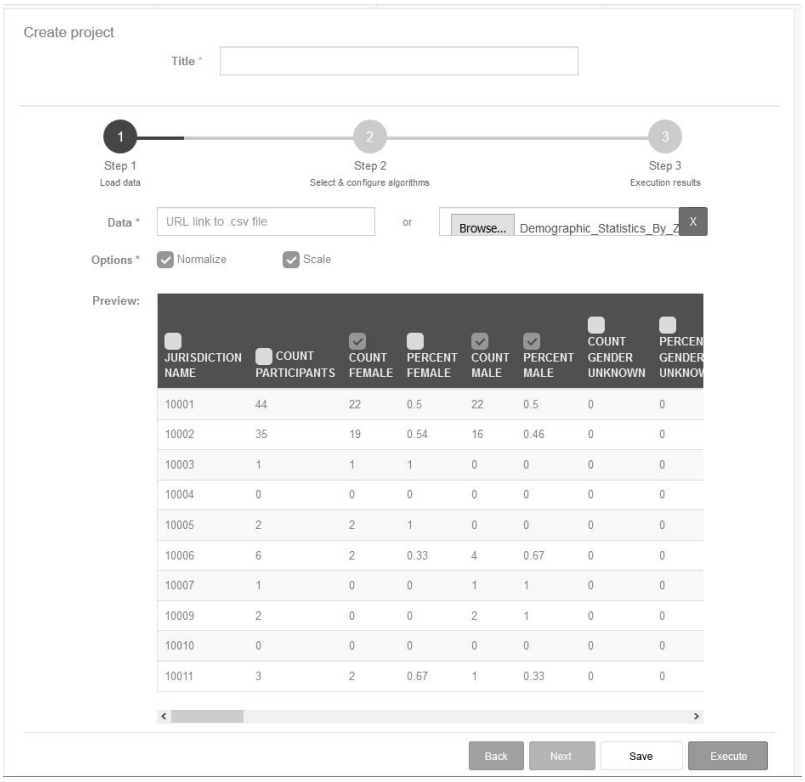


Figure 2. Project design. Step 1

After the input data loading first 10 lines are dynamically displayed for the demonstration. Meanwhile the ability to normalize into [0, 1] range and scale the data is provided along with the possibility to choose certain data features for the further analysis.

The field Title is used as a name of the project.

Next on the Step 2 (figure 3) the algorithm pipeline is built. REST-request provides the list of the supported libraries and available algorithms are displayed as a dropdown picklist.

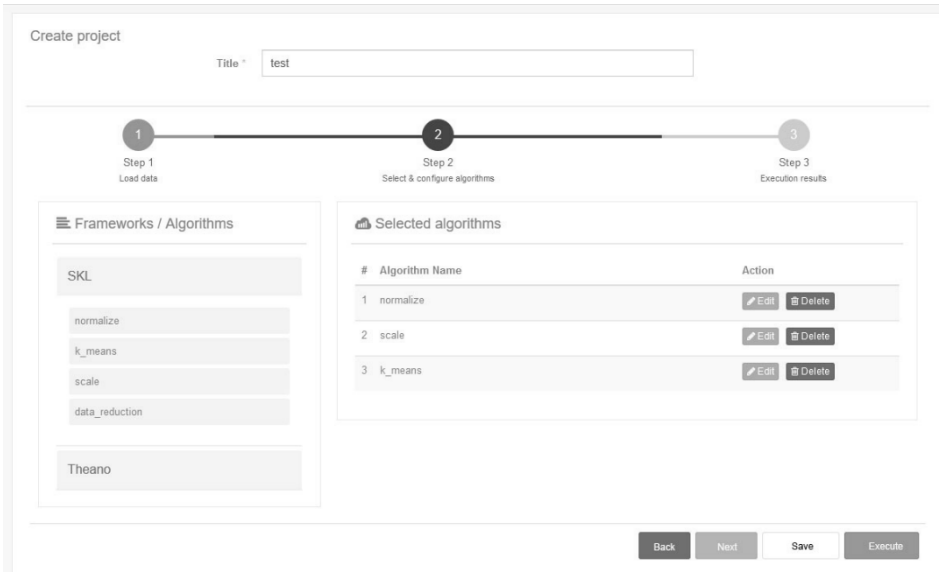


Figure 3. Project design. Step 2

The pipeline is formed by drag-and-drop of the algorithms to the Selected algorithms table. Execution order is determined by the order in the table, so if needed, the elements can be rearranged. The Edit button loads the parameters of the algorithm configuration (initially with default values).

If the created pipeline is not time consuming, on Step 3 (figure 4) user can immediately view the results. In case the result is not available instantly, it can be viewed afterwards via the list of projects.

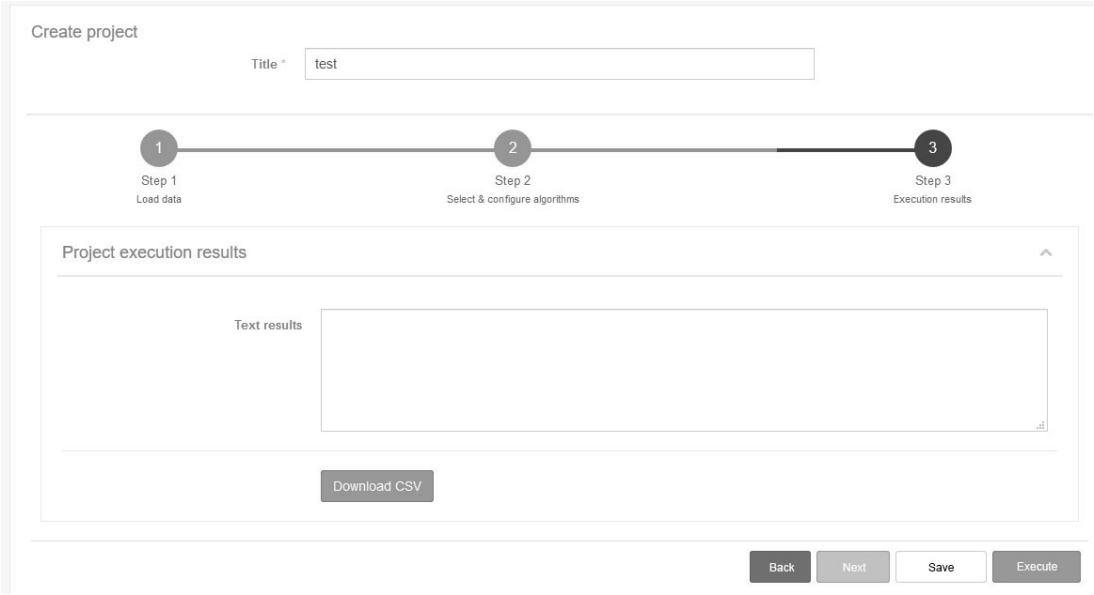


Figure 4. Project design. Step 3

Additionally, the user can save or execute the project again.

Image Captioning Example

The aspects of data analysis system for education and research application are discussed in [12-13]

As a showcase of the provided services application to the research problem let's consider the work done collaboratively by researchers of the ECM and IIT departments, BSUIR.

The case addresses the problem of image understanding. The problem can be stated as follows: given the image, provide the textual description of the situation depicted using simplified language constructions. In general, this problem is widely known as Image Captioning problem. The most successful approaches are, as a rule, hybrid architecture for object detection and semantic analysis [14]. Our solution similarly is based on deep convolutional networks and open semantic technology for intelligent systems design (OSTIS-technology) [15].

The architecture is built with the help of Apache Zeppelin project and the algorithm pipeline contains the following:

– image detection unit;

– graph construction;

– semantic analysis.

Image detection unit is built on the Faster R-CNN [16] architecture, internal algorithms for which are provided by TensorFlow via Apache Zeppelin. Detection results – object and regions – are passed to graph construction algorithm where possible subject-object relations are established. Next, semantic analysis unit, which is designed as an OSTIS-system, can determine image context as well as detect errors either in the object detection or in the graph. Depending on the error, «suspicious» regions are re-calculated (feedback loop) and using the corrected model subject-object relations are transformed into textual description. The analysis results can also be further integrated into knowledge base and be used to improve the semantic analysis.

The example of the object detection for the sample picture is shown in figure 5.

Figure 5. Detection results. Test image is provided by COCO dataset [17]

Based on the objects and regions initial closeness graph (figure 6.a) is built. Graph construction unit specifies the relations (figure 6.b). Semantic analysis unit determines the context – street, traffic – and detects possible errors (figure 6.c, dash-line). In figure 7, it is shown how re-detection

for the suspicious regions (by means of Mask R-CNN [18]) works. Corrected model is shown in figure 6.d.

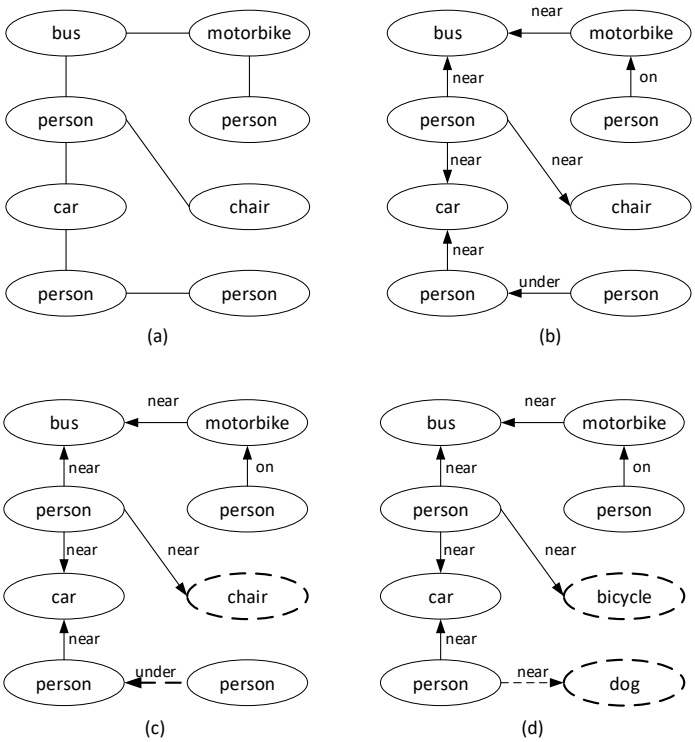
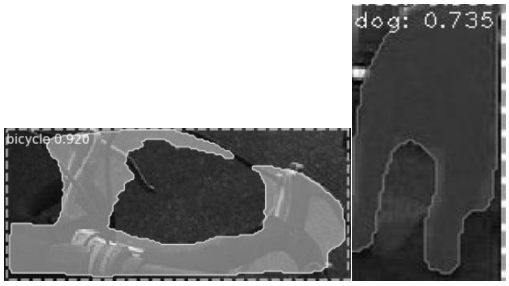


Figure 6. Graph correction steps



(a) (b)

Figure 7. Re-detection of the «suspicious» regions

Using final corrected model, by changing subject-object pairs into corresponding natural language constructions the following captions can be generated:

«**the** person **rides the** motorbike»

«**the** person **walks the** dog»

Conclusion

This work promotes a new modernized version of the BSUIR computing cluster, which enables new opportunities in large amounts of data analysis and processing, including image processing and semantic analysis.

References

[1] Demidchuk A. I. Uchebno-issledovatel'skaya sistema obrabotki bol'shikh dannykh / A. I. Demidchuk, D. Yu. Pertsev, D. I. Samal' // BIG DATA and Advanced Analytics. Minsk: BGUIR, 2017, pp. 170 – 173. (in Russ.)

[2] Sistema obrabotki bol'shikh dannykh na osnove vychislitel'nogo klastera BGUIR / D. I. Samal' [i dr.] // BIG DATA Advanced Analytics. Minsk: BGUIR, 2018, pp. 220 – 256. (in Russ.)

[3] Intellektual'naya obrabotka bol'shikh ob"emov dannykh na osnove tekhnologii MPI i CUDA. Laboratornyi praktikum : posobie / A. I. Demidchuk [i dr.]. Minsk : BGUIR, 2017, 60 p. (in Russ.)

[4] Zeppelin [Electronic resource]. URL: http://zeppelin.apache.org/ (access date: 20.01.2019).

[5] Scikit-learn: Machine Learning in Python [Electronic resource]. URL: https://scikit-learn.org/stable/ (access date: 20.01.2019).

[6] MLib Apache Spark [Electronic resource]. URL: https://spark.apache.org/mllib/ (access date: 20.01.2019).

[7] Theano 1.0.0 documentation [Electronic resource]. URL: http://deeplearning.net/software/theano/ (access date: 20.01.2019).

[8] Weka 3 – Data Mining with Open Source Machine Learning Software in Java [Electronic resource]. URL: https://www.cs.waikato.ac.nz/ml/weka/ (access date: 20.01.2019).

[9] Apache Spark – Unified Analytics Engine for Big Data [Electronic resource]. URL: https://spark.apache.org (access date: 20.01.2019).

[10] TensorFlow [Electronic resource]. URL: https://www.tensorflow.org (access date: 20.01.2019).

[11] JSON [Electronic resource]. URL: https://www.json.org (access date: 20.01.2019).

[12] Tatur M.M. Osobennosti postroeniya vychislitelei intellektual'noi obrabotki dannykh / M.M. Tatur // Informatika, №1(45), Minsk, 2015, pp. 39 – 44. (in Russ.)

[13] Tatur M.M., Iskra N.A. Intelligent Data Analysis: From Theory to Practice // Otkrytye semanticheskie tekhnologii proektirovaniya intellektual'nykh sistem. Minsk, BGUIR, 2018, pp. 171 – 175.

[14] Hossain, MD. Zakir et al. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Computing Surveys 51.6, 2019, pp. 1–36.

[15] V. Golenkov and N. Gulyakina, Proekt otkrytoi semanticheskoi tekhnologii komponentnogo proektirovaniya intellektual'nykh sistem. Chast' 2: Unifitsirovannye modeli proektirovaniya [project of open semantic technology of component design of intelligent systems. part 2: Unified design models], Ontologiya proektirovaniya [Ontology of design], no. 4, 2014, pp. 34–53. (in Russ.)

[16] Ren, Shaoqing et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39.6, 2017, pp. 1137–1149.

[17] COCO – Common Objects in Context [Electronic resource]. URL: http://cocodataset.org/ (access date: 20.01.2019).

[18] He, Kaiming et al. Mask R-CNN. IEEE International Conference on Computer Vision (ICCV), 2017.

ВЕБ-ПРИЛОЖЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ ОБРАЗОВАНИЯ И НАУЧНЫХ ИССЛЕДОВАНИЙ

Д.Ю. Перцев
Старший преподаватель кафедры ЭВС БГУИР

Н.А. Искра
Заместитель заведующего кафедрой по учебно-методической работе, старший преподаватель

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: pertsev@bsuir.by.

Аннотация. В данной статье представлена обновленная концепция использования вычислительного кластера БГУИР, построенная в соответствии с моделью частного облака и предоставляющего услуги интеллектуального анализа данных. Представлены результаты исследований в области аннотирования изображений с применение кластера.

Ключевые слова: Интеллектуальный анализ данных, Облачные вычисления, Частное облако, Кластер, Аннотирование изображений.