Knowledge Pit - A Data Challenge Platform*

Andrzej Janusz[1], Dominik Slezak[1,2], Sebastian Stawicki[1,2], and Mariusz Rosiak

[1] Institute of Mathematics, University of Warsaw,
Banacha 2, 02-097, Warsaw, Poland
[2] Infobright Inc., Poland
Krzywickiego 34, lok. 219, 02-078 Warsaw, Poland
{janusza,slezak,stawicki}@mimuw.edu.pl
mariusz.rosiak@gmail.com
http://www.dominikslezak.org

Abstract. Knowledge Pit (https://knowledgepit.fedcsis.org) is a web platform created to facilitate organization of data mining competitions. Its main aim is to stimulate collaborative research for solving practical problems related to real-life applications of predictive analysis and decision support systems. What makes Knowledge Pit different from other data challenge platforms is the fact that it is a non-commercial project focusing on a collaboration with international conferences. It promotes the idea of open research and encourages young researchers to involve in projects related to data science. The platform can also be used as a e-learning tool to support data mining courses and for defining interesting student projects. In this paper we discuss the architecture of Knowledge Pit and highlight its main functionalities. We also overview some of the already finished data challenges that were organized using our web platform.

Key words: data mining competitions, collaborative research, web platform, e-learning

1 Introduction

In this short paper we briefly describe a web platform, called Knowledge Pit, created in order to support organization of data mining competitions. On the one hand, this platform is appealing to members of the machine learning community for whom competitive challenges can be a source of new interesting research topics. Solving real-life complex problems can also be an attractive addition to academic courses for students who are interested in practical data mining. On the other hand, setting up a publicly available competition can be seen as a form of outsourcing the task to the community. This can be highly beneficial to the organizers who define the challenge, since it is an inexpensive way to solve the problem which they are investigating. Moreover, an open data mining competition can become a bridge between domain experts and data analysts. In a longer perspective, it may leverage a cooperation between industry and academic researchers.

* The authors are partially supported by Polish National Science Centre grant DEC-2012/05/B/ST6/03215 and by Polish National Centre for Research and Development grants PBS2/B9/20/2013 and O ROB/0010/03/001.

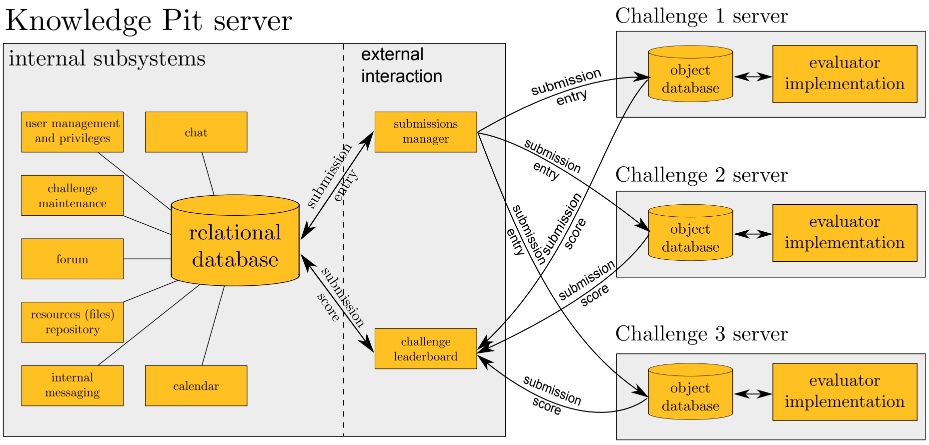

Fig. 1. A system architecture of the Knowledge Pit web platform.

2 System Architecture

The Knowledge Pit platform is designed in a modular way, on top of an open-source e-learning platform *Moodle.org* [1] and as such, it follows the best practices of a software development. The current modules of the platform include user accounts management system, competition management subsystems, time and calendar functionalities, communications features (i.e. forums and messaging subsystems), and a flexible interface for connecting automated evaluation services prepared to assess contestants' submissions.

Figure 1 shows an architecture schema of the Knowledge Pit platform. Its two main parts are the platform's engine located at a dedicated server and the evaluation subsystems. Currently, Knowledge Pit is hosted on a server belonging to Polish Information Processing Society (`http://pti.org.pl/`) and is located in the *fedcsis.org* domain.

The two main parts of the platform are the platform's engine and the evaluation subsystems. The first one provides interfaces for defining and maintaining of data challenges, management of user's profiles, submissions and private files, maintaining *Leaderboards*[3], and the internal messaging systems (competition forums, chats, as well as email and notification sending services). It is based on a very popular solution stack, i.e. Apache, MySQL and PHP [6, 9, 2]. Together they constitute a bridge between the platform and different groups of users (guests, participants of competitions, moderators and organizers of particular challenges, managers and administrators of the system).

The second part of the platform is responsible for assessment of solutions submitted by participants of particular competitions. Due to a flexible communication mechanism, this service may be distributed among several independent workstations, which guarantees the scalability of the evaluation process. Since evaluating submissions for

[3] A competition's Leaderboard is an on-line ranking of participants competing in that particular data challenge.

some competitions may require a lot of resources (e.g. memory, CPU time, disc I/O or database connections), this is a very important aspect of system's architecture. For example, the assessment of a single submission to AAIA'14 Data Mining Competition required constructing several Naive Bayes classification models for a data table consisting of $50,000$ objects and testing their performance on a different table with $50,000$ objects described by $11,852$ conditional attributes [3]. In that case, distribution of the required computations allowed for nearly real-time evaluation, even during the most busy moments of the competition.

Another advantage of separating the evaluation subsystems from the platform's engine is that it may be implemented in any suitable programming language, as a script or a stand alone compiled application that can use any external libraries. In this way, the responsibility for preparation of a suitable evaluation procedure can be delegated to organizers of individual competitions. In such a case, the only requirement for the implementation of the evaluator is that it should maintain a correct protocol of information exchange with the platform's engine. This flow of responsibilities frees Knowledge Pit from the things which it cannot cope with in a generic way. It also gives competition organizers a very flexible method of expressing their data mining task in a form of a fully customizable evaluation procedures. For instance, the evaluation procedure can be implemented in R language [8], in a form of a script that runs independently on several machines.

3 Examples of Data Challenges Hosted by Knowledge Pit

Knowledge Pit inaugurated in the beginning of 2014 and since then continues to organize successful data mining competitions in cooperation with international conferences. By June 2015 it had hosted 4 major competitions and a few local student projects. It currently has over 700 active users who participated in at least one data challenge and this number grows with every new competition. Below we list the recent competitions and shortly describe their scope. Typically, after completion of a contest, its overview and detailed descriptions of top solutions are published in proceedings of the associated conference.

3.1 AAIA'14 Data Mining Competition

AAIA'14 Data Mining Competition: Key risk factors for Polish State Fire Service[4] took place between February 3, 2014 and May 7, 2014. In this challenge the focus was on the feature selection problem and the data came from the public safety domain. We asked members of the machine learning community to identify characteristics extracted from the EWID reports [5], which are useful for predicting whether any people were harmed during a given incident.

[4] Web page: `https://knowledgepit.fedcsis.org/contest/view.php?id=83`

3.2 AAIA’15 Data Mining Competition

AAIA’15 Data Mining Competition: Tagging Firefighter Activities at a Fire Scene[5] took place between March 9, 2015 and June 5. It was a continuation of the contest initiated during the previous edition of the data challenge associated with International Symposium on Advances in Artificial Intelligence and Applications (the AAIA conference series) [3]. The topic was related to real-time screening of firefighters’ vital functions and monitoring of ongoing physical activities at the incident scene [7].

3.3 PAKDD’15 Data Mining Competition

PAKDD’15 Data Mining Competition: Gender Prediction Based on E-commerce Data[6] took place between March 23, 2015 and May 3 of the same year. The task in this competition was to reconstruct the information about user’s gender from product viewing logs from an on-line store. The data set was obtained from simulations of product viewing activities for user with known gender and was provided by FTP Group - the leading information and communication technology enterprise in Vietnam. The results of this competition were presented at a major Asia-Pacific data mining conference PAKDD’15 and were acclaimed by the industry representatives from FTP Group.

3.4 IJCRS’15 Data Challenge

IJCRS’15 Data Challenge: Mining Data from Coal Mines[7] started April 13, 2015 and lasted until June 25, 2015. The task was to come up with a prediction model which could be effectively applied to foresee warning levels of methane concentrations at three methane meters placed in a longwall of the mine [4]. The data used in the competition came from an active Polish coal mine. They consisted of multivariate time series corresponding to readings of sensors used for monitoring the safety conditions at the longwall.

4 Future Development of the Platform

In this paper we briefly described our data challenge platform called Knowledge Pit. It is worth noticing that this non-commercial project is far from complete. We are continuously searching for new topics of data mining contests related to important practical issues. We are also working on developing new features and functionalities for our platform. One example of such a feature is a support for an evaluation system that not only assesses submissions of participants with regard to their predictive quality but also tries to grasp adaptiveness of the proposed solutions, i.e. how fast they can produce results with sufficient quality and how much training data do they need.

[5] Web page: `https://knowledgepit.fedcsis.org/contest/view.php?id=106`

[6] Web page: `https://knowledgepit.fedcsis.org/contest/view.php?id=107`

[7] Web page: `https://knowledgepit.fedcsis.org/contest/view.php?id=109`

References

1. Cole, J.: Using Moodle. O'Reilly, first edn. (2005)
2. Isaacson, P.C.: Building a simple website using open source software (gnu/linux, apache, mysql, and python). J. Comput. Sci. Coll. 19(1), 286–288 (Oct 2003), `http://dl.acm.org/citation.cfm?id=948737.948777`
3. Janusz, A., Krasuski, A., Stawicki, S., Rosiak, M., Ślęzak, D., Nguyen, H.S.: Key Risk Factors for Polish State Fire Service: a Data Mining Competition at Knowledge Pit. In: Ganzha, M., Maciaszek, L.A., Paprzycki, M. (eds.) Proceedings of FedCSIS'14. pp. 345–354 (2014)
4. Janusz, A., Ślęzak, D., Sikora, M., Wróbel, Ł., Stawicki, S., Grzegorowski, M., Wojtas, P.: Mining data from coal mines: IJCRS'15 Data Challenge. In: Proceedings of IJCRS'15 (2015), in print November 2015
5. Krasuski, A., Janusz, A.: Semantic tagging of heterogeneous data: Labeling fire & rescue incidents with threats. In: Ganzha, M., Maciaszek, L.A., Paprzycki, M. (eds.) Proceedings of FedCSIS'13. pp. 295–302 (2013)
6. Lee, Ware, B.: Open Source Development with LAMP: Using Linux, Apache, MySQL and PHP. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2002)
7. Meina, M., Janusz, A., Rykaczewski, K., Ślęzak, D., Celmer, B., Krasuski, A.: Tagging firefighter activities at the emergency scene: Summary of aaiaŠ15 data mining competition at Knowledge Pit. In: Ganzha, M., Maciaszek, L.A., Paprzycki, M. (eds.) Proceedings of FedCSIS'15 (2015), in print September 2015
8. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008), `http://www.R-project.org`
9. Rosebrock, E., Filson, E.: Setting Up LAMP: Getting Linux, Apache, MySQL, and PHP Working Together. SYBEX Inc., Alameda, CA, USA (2004)