Areca: Online Comparison of Research Results

David Urbansky, Klemens Muthmann, Lars Kreisz, Alexander Schill
Dresden University of Technology
01062 Dresden
Germany
{david.urbansky, klemens.muthmann, lars.kreisz, alexander.schill}@tu-dresden.de

Abstract

To experiment properly, scientists from many research areas need large sets of real world data. Information retrieval scientists for example often need to evaluate their algorithms on a dataset or a gold standard. The availability of these datasets often is insufficient and authors with the same goal do not evaluate their approaches on the same data. To make research results more transparent and comparable, we introduce Areca, an online portal for sharing datasets and/or the results that were reached with the author's algorithms on these datasets. Having such an online comparison makes it easier to grasp the state-of-the-art on certain tasks and drive research to improve the results.

Problem Statement

When a new algorithm or technique is developed to solve a certain problem, this new approach needs to be evaluated and compared to the state-of-the-art in that field. For very common problems, such as text classification, named entity recognition, and face detection, a number of datasets have been used by different researchers in order to allow them to compare their work directly. However, these datasets do not exist for every task, forcing researchers to develop their own datasets to evaluate their algorithms. Many times these datasets are even made public and the dataset location is mentioned in the paper so that other researchers can use it, too. However, these hosted datasets often disappear from university servers and can not be accessed by others anymore as soon as the paper gets older.

There are two problems that need to be addressed:

- We need a centralized point to store and find reusable datasets for different research domains.
- We need to be able to compare our computed results on the dataset and post them online for other researchers to directly compare with them.

Related Work

To the best of our knowledge there is no system which tries to solve the two problems we are addressing with Areca. There are, however, online repositories for datasets. Pangaea (Diepenbroek et al. 1993) is an information system for geoscientific and environmental data. Every individual researcher or project can publish their data under the Creative Commons license on the platform. The system is, however, limited to the geo domain and only hosts datasets, without any evaluation results. The UCI Machine Learning Repository (Frank and Asuncion 2010) is a repository of almost 200 datasets that can be used to evaluate machine learning algorithms and intelligent systems. For each dataset, information about the content, the relevant papers, and suitable tasks is stored. The UCI repository partially solves the problem of dataset reuse for the machine learning community. Other communities do not always have similarly advanced repositories. Also, there is no comparison of algorithms for different tasks on the datasets, that is, the problem about comparison remains unsolved. Other repositories such as Infochimps[1] and Datamob[2] are domain independent, that is, they host datasets from different fields and store information, such as tags, the license, and a description. The intention of these repositories is, however, to make datasets more accessible, not to compare research results on these datasets. MyExperiment (Goble and De Roure 2007) is a platform where researchers can share scientific workflows and research objects. It does, however, not allow comparison between algorithms on shared tasks. Research challenges as the ones held by TREC[3] or DARPA[4] give further evidence on the importance of comparable research. Those competitions are, however, rare and often focused on a small set of tasks.

Solution

With Areca[5] we try to solve the previously mentioned problems. That is, (1) we created a centralized platform to store datasets and (2) we provide a point where researchers can compare their results on the same datasets. The system provides the following functionality:

- Login via OpenID (also Google and Yahoo Login) to avoid that users need to register.

[1] http://infochimps.com/
[2] http://datamob.org/
[3] http://trec.nist.gov
[4] http://www.darpa.mil
[5] http://areca.co

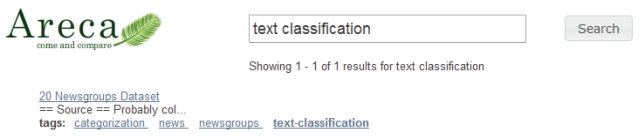

Figure 1: Search for a Dataset or Result.

- Fulltext search of dataset descriptions and tags.
- Upload and edit a dataset. The author can either upload the dataset directly from their disk, upload it by providing a URL, or simply referring to a dataset by linking it. The author can provide a description of the dataset in Wiki notation, add tags, and choose a license.
- Create and edit tasks. On a single dataset there could be different tasks that are of interest. For example, on a dataset of English newspaper articles, the task could be to classify the articles by topic or extract named entities.
- Add results to tasks. The main innovation of Areca is that authors can now post their results on the created tasks. For each result a number of traits can be created. For example, an author could post what their classification accuracy (trait) on the classification task of the newspaper dataset is.
- Visualize results on tasks. Results from all authors on a given task are visualized and compared against each other on each defined trait.
- Each result can link to a paper explaining the methods in detail. The paper information does not have to be pasted into the system when the author is a member of the research community Mendeley. In that case the information can directly be retrieved from the Mendeley API[6].

Scenario

Let us consider the following scenario. The researcher Alice has an idea for an algorithm for text classification. Before she starts experimenting, Alice needs to know which algorithms are the best so far and, therefore, what her baseline should be. She would need to read the most recent papers about text classification and hope that they compared their algorithms with the current state-of-the-art. The next step would be finding datasets on which she could run her algorithms on. Again, papers by other researchers might give her hints about what can be used and where it can be found. After finally finding the latest research results, the description of the algorithms, and hopefully some datasets, she can start comparing her approach to existing ones manually. Alice now may publish her findings, hoping that other researchers will find her results.

The same process can be simplified with Areca. The search starts with "text classification" on the platform as shown in Figure 1. Alice now sees all datasets that were used for the task and all the evaluations of the algorithms. Figure 2 shows the comparison page of different algorithms on one task and dataset[7]. On this page, Alice determines which algorithm performs best for a task, who the authors of the result are,

[6]http://dev.mendeley.com/

[7]http://areca.co/2/20-Newsgroups-Dataset/Text-Classification

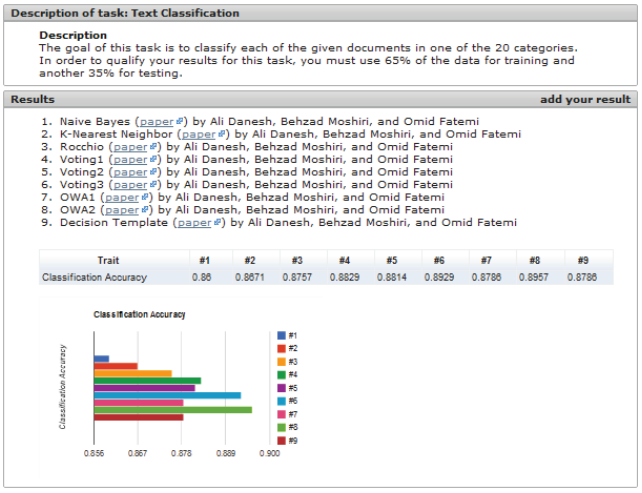

Figure 2: Comparing Text Classification Algorithms on a Dataset.

and in which paper she finds more about the approach. She now runs her own algorithm on the dataset following the instructions of the task and publish her results on the same page as well. Furthermore, when publishing her results in a paper, she points to this result and comparison page. Alice does not have to host the dataset and/or her results on her own web server that might be inaccessible in the near future.

Conclusion and Outlook

In this paper, we introduced Areca, an online repository for research datasets and evaluation results. Our goal is to make research results more comparable and make it easier to find state-of-the-art algorithms on research problems. Areca is work in progress. The next step is to ensure that researchers publish correct results. Since everybody can post basically anything, we need a trust system that researchers can use. One idea is to use the community and give them the ability to "distrust a result". If too many researchers do not trust a result it will be removed unless the proper evidence is shown. Another idea we're looking into is having multiple instances of Areca so that each research group can have its own that can be synchronized to the global instance.

References

Diepenbroek, M.; Grobe, H.; Schindler, U.; and Sieger, R. 1993. Publishing network for geoscientific environmental data. http://www.pangaea.de/.

Frank, A., and Asuncion, A. 2010. UCI machine learning repository. http://archive.ics.uci.edu/ml.

Goble, C., and De Roure, D. 2007. myExperiment: social networking for workflow-using e-scientists. In *Proceedings of the 2nd workshop on Workflows in support of large-scale science*, 1–2. ACM.