Scientific report of the meeting Evaluation as a Service (EaaS), funded by ELIAS

Summary

In this text, we summarize the outcome of the ``Evaluation-as-a-Service'' workshop that was held on the 5th and 6th March 2015 in Sierre, Switzerland, funded by the ELIAS project of the ESF (European Science Foundation). The objective of the meeting was to bring together initiatives that use cloud infrastructures, virtual machines, APIs (Application Programming Interface) and related projects that provide evaluation of information retrieval or machine learning tools as a service (EaaS).
The standard approach to evaluating Information Retrieval (IR) systems involves distributing the data to the groups developing the systems so that they perform the evaluation locally. However, this approach of distributing data is often not practical, as the data may be:

- Huge – In order to obtain realistic evaluation results for IR, the evaluation should be done on realistic amounts of data. In the case of web search, this could be Petabytes of data. The current common approach of sending this data on hard disks through the postal service or via download has its limitations.
- Non-distributable – In many cases, it is not permitted to distribute data due to privacy, terms of service, or commercial sensitivity of the data. Privacy is the major concern for patient records. Even though law permits the distribution of anonymized medical records, large-scale anonymization can only be accomplished automatically, which data owners usually do not trust. For example, the Twitter terms of service forbid redistribution of tweets, while query logs are not made available for researchers after the AOL debacle in 2006. Distribution of company documents for the evaluation of enterprise search would not be permitted due to the commercial sensitivity of the data.
- Real-time – Companies working on real-time systems, such as recommender systems, are often not interested in evaluation results obtained on static historical data, in particular if these data have to be anonymised to allow distribution, as these results are too far removed from their operative requirements.

A number of initiatives are currently working to solve the above challenges. These initiatives all basically implement Evaluation-as-a-Service (EaaS), either making available APIs to access the data in a controlled way, or Virtual Machines (VMs) on which systems should be deployed. In order to organize these evaluation services, various aspects need to be considered. An overview of these aspects as developed during the workshop is given in Figure 1.

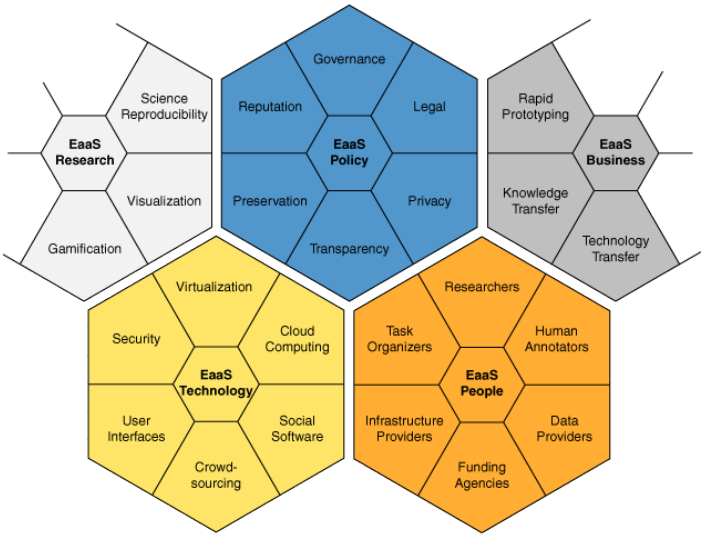

Figure 1: Many aspects that need to be taken into account for the Evaluation as a Service paradigm.

After having presented current initiatives and experiences of these initiatives, including the personal background of all workshop participants, the planning of the group work for the second day was done in details.
At the end of the meeting a SIGIR forum paper was outlined and the main points for a white paper as follow up of the workshop was also written. This white paper is planned to be written by summer 2015.
A mailing list and web page[1] were created for the group and the topic is expected to increase further in importance.

[1] http://eaas.cc/

Scientific content of the meeting

The workshop was dealing with the evaluation of information retrieval and information analysis systems in the widest sense and with providing such **Evaluation as a Service** (EaaS). This implies the use of APIs and or cloud infrastructures instead of standard approaches where data and query topics are distributed and then results are compared based on submitted runs. Using a central infrastructure such as a cloud or local servers with virtual machines has the advantage to move the algorithms to the data instead of doing it the other way around [1]. This meeting was co-funded by the VISCERAL[2] project that has developed a cloud-based evaluation approach for medical image analysis and retrieval.

Motivation

The standard approach to evaluating Information Retrieval (IR) systems involves distributing the data to the groups developing the systems so that they perform the evaluation locally. However, this approach of distributing data is often not practical, as the data may be:

- **Huge** – In order to obtain realistic evaluation results for IR, the evaluation should be done on realistic amounts of data. In the case of web search, this could be Petabytes of data. The current common approach of sending this data on hard disks through the postal service or via download has its limitations.
- **Non-distributable** – In many cases, it is not permitted to distribute data due to privacy, terms of service, or commercial sensitivity of the data. Privacy is the major concern for patient records. Even though law permits the distribution of anonymized medical records, large-scale anonymization can only be accomplished automatically, which data owners usually do not trust. For example, the Twitter terms of service forbid redistribution of tweets, while query logs are not made available for researchers after the AOL debacle in 2006. Distribution of company documents for the evaluation of enterprise search would not be permitted due to the commercial sensitivity of the data.
- **Real-time** – Companies working on real-time systems, such as recommender systems, are often not interested in evaluation results obtained on static historical data [3], in particular if these data have to be anonymized to allow distribution, as these results are too far removed from their operative requirements. Continuous evaluation infrastructures could leverage this problem.

In addition to the abovementioned challenges there are many questions regarding reproducibility of scientific results [4] and often it is said that a scientific paper in computer science should include the data used for evaluation and the executable used for obtaining the results. By making data sets citable and keeping them available in connection with VMs of the data analysis tools would also correspond to these criteria as the results can actually be reproduced easily [2].

Another challenge in machine learning research is that as soon as test data are available people tend to optimize solutions with at least partly using the test data, even if the ground truth is maybe not available. Giving no access to the test data to the participants at all, only allowing access for algorithms could again avoid this problem and really make results comparable.

Objectives

The objectives of the workshop were to bring together people working on evaluation as a service of similar initiatives in data analysis or information retrieval. By **sharing experiences** on the various approaches it should be possible to work out the advantages, inconveniences and also differences based on the user groups and the objectives of the initiatives.

Another objective was to **reach a wider audience** with the preparation of a white paper on the outcomes of the workshop to maximize impact. We felt that this is a topic of potentially large impact in terms of practical implications, as big data is a hype topic and evaluation on large data sets is still relatively rare as distribution of data is non-trivial.

A third objective was to **create a community** around the topic to have an interest group and other person to ask new questions and discuss best practices in the field. This community can then also respond to outside requests, present the ideas and potentially be involved in common research project proposals.

Used techniques

Persons were invited based on their experience with the topic at hand and the objective was to have one person per initiative and a large coverage. Participation was from the start aimed to be international covering the initiatives we knew and we found when searching the web. A request of the ELIAS organizers was to include infrastructure provides and we contacted the proposed persons of Microsoft and Yandex but none of them was able to come. One

[2] http://visceral.eu/

person formerly working at Microsoft for evaluation initiatives was present and was able to take the commercial perspectives into account as well.
The following people participated at the meeting representing a variety of initiatives; often persons presented also other evaluation initiatives than those mentioned to take into account other aspects as well:

- TREC Microblog (Jimmy Lin)
- TIRA (Martin Potthast, Tim Gollub)
- BioASQ (Anastasia Krithara)
- VISCERAL (Allan Hanbury, Henning Müller, Ivan Eggel)
- CLEF Newsreel (Frank Hopfgartner)
- CodaLab (Simon Mercer)
- C-BIBOP (Jayashree Kalpathy-Cramer)
- Living Labs (Krisztian Balog)
- NTCIR initiatives (Noriko Kando)

Several other initiatives were contacted, such as Mirex, which has much experience in music retrieval where participants submit algorithms that analyze data and are executed in the same environment. Not all of them were able to join with the short notice given after acceptance of the ELIAS funding request.
In terms of used techniques at the meeting we started with **presentation of existing initiatives** in standard presentations during two sessions and then continued with **thematic discussions** on advantages and inconveniences of the approaches and the techniques used and which other initiatives existed related to the work discussed. At the end of the day group sessions were prepared based on the outcomes. **Group work** was then done on the second day before the group discussions were presented to all participants and discussed.
Then the structure of the white paper was prepared

Results
All initiatives present at the workshop were described in slides by the participants to get an overview of the types of challenges, the communities behind the data analysis and the solutions chosen. In addition the positive and negative experiences were listed by the participants to collect challenges and constraints for the further discussions.
We also identified several initiatives that were not present but might be very interested in the outcomes or could give additional feedback:

- Mirex music retrieval community that has used executables to be run on protected music pieces;
- Sage Synapse biomedical network, on biomedical data, machine learning;
- myExperiment, Taverna is more a workflow engine but shares some common goals;
- ChaLearn runs competitions and has different types of infrastructures;
- Yahoo pipes;
- Recomputation.org;
- (VideoBrowser Showdown[3]) allows submitting search results in a competitive evaluation;
- Kaggle[4] as a commercial system and provider, many participants
- Delve
- EvaluatIR (http://wice.csse.unimelb.edu.au:15000/evalweb/ireval/)
- OpenML
- MLcomp
- SEP (http://sepwww.stanford.edu/doku.php?id=sep:about:about)
- TunedIT (http://www.tunedit.org/)
- 3X (http://netj.github.io/3x/)
- runmycode (http://www.runmycode.org/)

This list is surely not complete can does give hints and ideas for other experiences in related fields.
Stakeholders were also identified in a systematic way as:

- task organizers;
- researchers (participants);
- data annotators or task developers;
- data providers;
- funding agencies;
- and infrastructure providers.

[3] http://www.videobrowsershowdown.org/
[4] http://www.kaggle.com/

Each stakeholder can have a different viewpoint on the aspects of an evaluation task and different interest but they are all linked by a common evaluation environment such as a cloud in our case.
In terms of large domains five main areas were identified and discussed (see also figure 1):

- **Policy** definition is necessary to push towards solutions such as evaluation as a service as it has many advantages for funding organizations and other stakeholder. Aspects that need to be taken into account for the general governance are transparency, privacy of data, reputation and related aspects.
- In terms of **research** there are clear advantages on reproducibility but also other aspects that could be taken into account.
- **People** involved in evaluation activities include all those just listed above.
- The **business** in terms of EaaS can include people providing data and challenges and estimating a good outcome, so rapid prototyping but there are also interesting aspects for technology transfer and knowledge transfer.
- In terms of **techniques** there is also a variety ranging from virtualization to cloud aspects, security requirements, social computing, and also crowdsourcing for the relevance judgments on the data.

Four pages are not sufficient to summarize all outcomes of the text and the foundations laid for next steps. We plan that the white paper as a follow up will go deeper into several of the discussed aspects and will also open up new research directions.

Conclusions

Based on the feedback from the participants and our own impression we feel that the workshop was a big success! It brought people together that would have never been able to discuss these topics as deeply in any other environment and without the provided funding. Such small expert workshops can be extremely useful to elaborate on new ideas and directions if knowledge is distributed geographically and scarce. The format including presentations, time for in-depth discussions and group work also seems to correspond well to the objectives we had before the workshop.

References

[1] Allan Hanbury, Henning Müller, Georg Langs, Marc André Weber, Bjoern H. Menze, and Tomas Salas Fernandez. Bringing the algorithms to the data: cloud-based benchmarking for medical image analysis. In CLEF'12: Proceedings of the 3rd International Conference of the CLEF Initiative, pages 24–29. Springer Verlag, 2012.
[2] Tim Gollub, Benno Stein, and Steven Burrows. Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In SIGIR'12: Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval, pages 1125–1126. ACM, 2012.
[3] Frank Hopfgartner, Benjamin Kille, Andreas Lommatzsch, Torben Brodt, and Tobias Heintz. Benchmarking News Recommendations in a Living Lab. In CLEF'14: Proceedings of the 5th International Conference of the CLEF Initiative, pages 250–267. Springer Verlag, 2014.
[4] Jinfeng Rao, Jimmy Lin, and Miles Efron. Reproducible experiments on lexical and temporal feedback for tweet search. In ECIR'15: Proceedings of the 37th European Conference on Information Retrieval, pages 755–767, Vienna, Austria, 2015.

Assessment of the results & impact of the event on the future directions of the field

The content of the meeting was extremely stimulating for all participants and the feedback of everyone was very positive on the lively discussions and the possibility to exchange with other people working on related topics.
Few people have actually worked on these or similar challenges used cloud approaches for benchmarking but all participants agreed that this will strongly rise in the coming years, as these problems currently block scientific advances in many domains where large-scale data are need and where data are potentially confidential or difficult to share. Being able to exchange practical experiences with other partners will have a lasting experience as will have the creation of a community around this topic.
Based on the first discussion a paper for the SIGIR forum with the main outcomes was written and submitted. It should be published in June, also with the goal to well communicate the discussions at the workshop and the main outcomes. The goal is not to create a closed community but be open and stimulate discussion on the topic.
To have a more lasting impact a white paper was prepared with specific sections already during the workshop and parts of the work were assigned to participants. Such a white paper should allow for a deeper analysis of the expert workshop and the experiences gained in several projects with Evaluation as a Service. The international participation form not only many European countries but also the USA and Japan allowed to combined views of several scientific organizations and allow global views as it seems logic in terms of such a large scale scientific approach.
As concrete steps a web page was reserved and initial content added[5]. For the workshop participants a mailing list was set up to also ease communication and also to make extensions of the ideas easier.
Among European partners concrete ideas on submitting EU funded research were discussed and these may still depend on the upcoming calls for projects.
All in all and based on the feedback from the participants the workshop was a big success. It brought together people from different domains who only partly knew each other but realized that they shared several common challenges and ideas for approaches to solve them. We feel that the outcomes in terms of a community but also regarding a potential impact on scientific evaluation can be important and we would like to thank ELIAS, as without the support this meeting would have been impossible.

[5] http://eaas.cc/

Annex 1: programme of the meeting

The meeting schedule was planned for the two days with fixed breaks and clear topics. The plan was in the end taken in a flexible manner to attribute much time to discussions and allow for changes based on the presentations and results in the discussions. The meeting took place in Sierre, Switzerland at the main building of the HES-SO in Sierre at Route de la Plaine. All participants were staying in the same hotel in Salquenen, around 4km from Sierre and the first evening a common dinner was taking place there and the second day the common dinner took place in Sierre.

Day0:
20h00: Common dinner for all workshop participants

Day1:
9h00-10h30: Existing initiatives using cloud-based evaluation and presentation of each participant
- VISCERAL, PAN, TREC microblog, BioAsq

10h30-11h00 coffee break
11h00-12h30: Existing initiatives using cloud-based evaluation and presentation of each participant
- CodaLab, NTCIR, CLEF newsreel, LivingLabs, C-BIBOP

12h30-14h00: lunch break

14h00-15h30: Summarizing the outcomes of the presentations of the initiatives
- technical infrastructures for storage and computation (public cloud, local servers)
- APIs and other approaches to distribute data

15h30-16h00: coffee break
16h00-17h30: Experiences on positive and negative aspects of the initiatives
- for organizers, participants and also data providers

17h30-18h30: Preparation of the hands on session for day 2 focusing on three topics
- preparation of groups for day 2

19h30: common dinner

Day2:
9h00-10h30: work in three groups:
- technical aspects,
- regulatory (political) aspects,
- emotional aspects.

10h30-11h00 coffee break
11h00-12h30: presentation of the group work to all participants,
- work on the black board and Google Docs to not forget any aspects

12h30-14h00: lunch break

14h00-15h30: structuring of the content of the three groups with implications & structure of the white paper
15h30-16h00: coffee break
16h00-17h30: attribution of work for
- the sections of the white paper,
- the SIGIR forum paper,
- the mailing list,
- the web page.

Annex 2: full list of speakers and participants

PD. Dr**. Allan Hanbury**, male
Initiative: VISCERAL
Vienna University of Technology
Institute of Software Technology and Interactive Systems
Information & Software Engineering Group
Favoritenstrasse 9-11/188, A-1040 Vienna, Austria
Tel. +43 1 58801 188310
hanbury@ifs.tuwien.ac.at

Prof. Dr. **Henning Müller**, male
Initiative: MultimediaEval, VISCERAL
HES-SO Valais
Techno-Pôle 3, 3960 Sierre, Switzerland
Tel. +41 27 606 9036
henning.mueller@hevs.ch

Ivan Eggel, male
Initiative: VISCERAL and research infrastructures project
HES-SO Valais
Techno-Pôle 3, 3960 Sierre, Switzerland
Tel. +41 27 606 9036
ivan.eggel@hevs.ch

Dr. **Frank Hopfgartner**, male
Initiative: TREC Newsreel
University of Glasgow
1 University Gardens
Room 205
Glasgow G12 8QQ, United Kingdom
Tel. +44 141 330 2472
frank.hopfgartner@glasgow.ac.uk

Prof. Dr. **Krisztian Balog**, male
Initiative: Living Labs,
University of Stavanger
Dept. of Electrical Engineering and Computer Science
NO-4036 Stavanger, Norway
Tel. +47 51 83 17 88
krisztian.balog@uis.no

Dr. **Simon Mercer**, male
Initiative: Codalab, former Microsoft employee responsible for CodaLab open source development
Currently no professional address available
Formerly: Director health and wellbeing, Microsoft research connections
simonm@ihmail.com

Prof. Dr. **Noriko Kando**, female
Initiative: NTCIR
Information and Society Research Division
National Institute of Informatics (NII)
Rm.1507, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, JAPAN
Tel. +81-3-4212-2529
Noriko.Kando@nii.ac.jp

Prof. Dr. **Jayashree Kalpathy-Cramer**, female
Initiative: C-BIBOP
Martinos Center for Biomedical imaging
Harvard Medical School
Building 149, Room 2301 13th Street
Charlestown, MA 02129 USA
Tel. +1 617 724 4657
Kalpathy@nmr.mgh.harvard.edu

Dr. **Martin Potthast**, male
Initiative: CLEF PAN
Bauhaus-Universität Weimar
Digital Bauhaus Lab
Bauhausstraße 9a
99423 Weimar, Germany
Tel. +49 3643 58 3567
martin.potthast@uni-weimar.de

Dr. **Tim Gollub**, male
Initiative: CLEF PAN
Bauhaus-Universität Weimar
Bauhausstraße 9a
99423 Weimar, Germany
Tel. +49 3643 58 3566
tim.gollub@uni-weimar.de

Dr. **Anastasia Krithara**, female
Initiative: BioAsq
Software & Knowledge Engineering Laboratory (SKEL)
Institute of Informatics & Telecommunications (IIT)
National Centre for Scientific Research "Demokritos" (NCSR)
Patriarchou Grigoriou and Neapoleos, Agia Paraskeui, Athens, Greece
Tel: (+30) 210-6503172
akrithara@iit.demokritos.gr

Prof. Dr**. Jimmy Lin**, male
Initiative: TREC Microblog
Hornbake Building, South Wing
The iSchool — College of Information Studies
University of Maryland
College Park, MD 20742, USA
Tel. +1 301 314-9145
jimmylin@umd.edu