Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service

Tim Gollub, Benno Stein, Steven Burrows
Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

With its close ties to the Web, the IR community is destined to leverage the dissemination and collaboration capabilities that the Web provides today. Especially with the advent of the software as a service principle, an IR community is conceivable that publishes experiments executable by anyone over the Web. A review of recent SIGIR papers shows that we are far away from this vision of collaboration. The benefits of publishing IR experiments as a service are striking for the community as a whole, and include potential to boost research profiles and reputation. However, the additional work must be kept to a minimum and sensitive data must be kept private for this paradigm to become an accepted practice. To foster experiments as a service in IR, we present a Web framework for experiments that addresses the outlined challenges and possesses a unique set of compelling features in comparison to existing solutions. We also describe how our reference implementation is already used officially as an evaluation platform for an established international plagiarism detection competition.

Categories and Subject Descriptors: H.5.3 [Information Systems]: Information Interfaces and Presentation—Group and Organization Interfaces
Keywords: Open Evaluation, Experiment Management, Result Dissemination

1. MOTIVATION AND SURVEY

Within IR research, the integration of previous work in one's experiments is significantly simplified if the data and software assets are published together with the papers. In this respect, Armstrong et al. [1] pointed out the verification problems that arise if research assets are not streamlined with the latest publications. In our view, the most convenient way to enhance comparability in experiments is to publish experiments as an online service where researchers can verify experimental results and explore alternate parameter settings.

To explore how IR research is published today, we reviewed all 108 full papers from the SIGIR 2011 proceedings concerning experiment assets. The results in Figure 1 show the extent to which the authors published their *data*, their *software*, and whether the experiments were hosted as a *service*. All three aspects need to be addressed to make publishing data and software more widespread. The questions we posed when examining the proceedings and our findings are given as follows.

Data. *Are any of the datasets used in the research publicly available?* We observe that 51% of the papers use a publicly available dataset for experiments. In this respect, evaluation initiatives such as TREC do a great job in supplying researchers with open datasets. The papers from the Image Search, Indexing, Retrieval Models, Test Collections, and Web Queries sessions all used public datasets. Other datasets such as search engine query logs require *non-disclosure* due to commercial and sensitivity reasons, which accounts for much of the unavailability of data. Correspondingly, only two of eight Query Analysis papers used public datasets, which was second lowest and only ahead of the Summarization papers, where no public datasets were used.

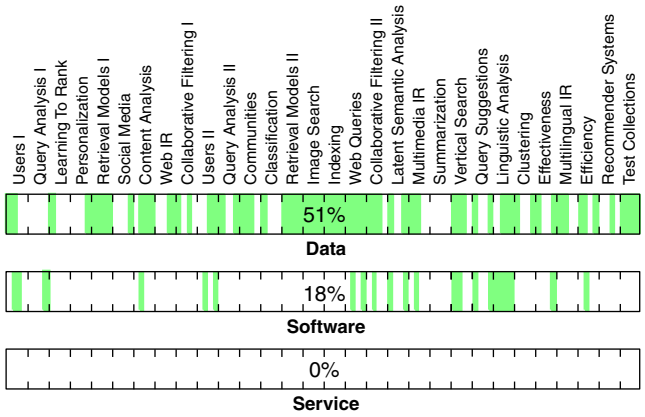

Figure 1: Assessment of the 108 full papers at SIGIR 2011 with respect to the publication of data, software, and experiments hosted as a service. The papers are ordered as they appear in the conference proceedings, and the labels on the horizontal axis denote the 30 sessions. The gray bars denote papers that meet the respective criteria.

Software. *Are any of the software assets used in the research publicly available?* We observe that 18% of the papers use shared software contributions, and hence publishing software is currently a minority practice in IR research.[1] Only the Linguistic Analysis papers published software with their papers in all cases. One explanation for this low ratio is the *lack of acknowledgment* researchers receive for publishing software. To foster software release, new incentives must be created that turn invested time in developing reusable software into reputation gain for the researchers.

Service. *Are any of the experiments in the research provided as an online service?* We observe that providing experiment software as a service is not practiced at all in the examined papers. The application closest to a web service is the online demonstration of ViewSer [2]. The lack of a *compelling web framework* that researchers can use to easily transform their experiments into a web service is a plausible explanation.

[1] A further 17% of researchers instead published algorithms in their papers as a compromise.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. PROPOSED FRAMEWORK

Motivated by the observations from the paper study above, we propose the development of an online framework to foster publishable IR experiments. The proposal is based on the needs for local installation, web dissemination, platform independence, result retrieval, and peer to peer collaboration. Our assessment of existing experimentation frameworks with respect to these goals is depicted in Table 1, which shows that none of the systems fully comply.

Table 1: Assessment of existing experimentation frameworks with respect to our five proposed design goals.

Tool	URL	Domain	1	2	3	4	5
evaluatIR	[1]	IR	×	✓	✓	✓	×
expDB	[2]	ML	×	×	×	✓	×
MLComp	[3]	ML	×	✓	×	✓	×
myExperiment	[4]	any	×	✓	✓	✓	×
NEMA	[5]	IR	×	✓	×	✓	×
TunedIT	[6]	ML, DM	✓	✓	×	✓	×
Yahoo Pipes	[7]	Web	×	✓	×	×	×

[1] http://www.evaluatir.org/
[2] http://expdb.cs.kuleuven.be/expdb/
[3] http://www.mlcomp.org/
[4] http://www.myexperiment.org/
[5] http://www.music-ir.org/
[6] http://www.tunedit.org/
[7] http://pipes.yahoo.com/

1. *Local Instantiation.* If data must be kept confidential, the framework must be able to reside with the data, hence the framework must be locally installable. Unlike centralized experiment platforms like MLComp and myExperiment, local instantiation allows experiments on sensitive data to be published as a service from a local host. External researchers can then use the service for comparison and evaluation of their own research hypotheses, whilst the experiment provider is in full control of the experiment resources.

2. *Web Dissemination.* URLs are definitive identifiers for digital resources. If all runs of an experiment are accessible over a unique URL, researchers can conveniently link the results in a paper with the experiment service used to produce them. Especially for standard pre-processing tasks or evaluations on private data, such a web service can become a frequently cited resource. In addition, attention can be attracted to one's work through integration of the service into home pages and blog articles. To address the issue of digital preservation, URLs should encode all information needed to recompute a resource, such as program and input parameter specifications, in case stored data is lost.

3. *Platform Independence.* The sophisticated and varying software and hardware requirements of information retrieval experiments as well as individual coding preferences of software developers render any development constraints imposed by the web framework critical for its success. Ideally, software developers can deploy experiments as a service unconstrained by the utilized operating system, parallelization paradigm, programming language, or data formats. Local instantiation is one key to realize this goal. Furthermore, the web framework must operate as a layer strictly on top of the experiment software and should use, instead of close intra-process communication such as in TunedIT, standard inter-process communication on the POSIX level and the file system to exchange information. This way, any running software can be deployed as a web service without internal modifications.

4. *Result Retrieval.* Especially for computationally expensive retrieval tasks, the maintenance of a public result repository can become a valuable asset of a research group. For example, experiment services that can index datasets with state-of-the-art natural language processing technology have the potential to raise the comparability of retrieval model research to a higher level. For clustering and result diversification research, comparability is enhanced by establishing static snapshots of the search results from major search engines regularly. The persistent storage of experiment results by the web framework is key to achieve this goal. Even if the public release of an experiment service is not desired, the framework is still useful if it assumes responsibility for managing the raw experiment results and making them available across a research team.

5. *Peer to Peer Collaboration.* Consider a scenario where a consortium of service providers become renowned *gatekeepers* for various streams of research, and maintain the community-wide repository of state-of-the-art algorithms, datasets, and experiment results on their web site. The gatekeepers drive the standardization of data formats and can, by utilizing the retrieval facility, stage competitions in a semi-automated fashion. A mechanism for connecting the local framework instances to a network of experimentation nodes has to be provided to achieve this scenario. Note that currently none of the experimentation platforms implements peer to peer collaboration.

3. REFERENCE IMPLEMENTATION

The reference implementation[2] of our proposal has been developed as a RESTful J2EE servlet. This implementation is also the official training and evaluation platform for the detailed comparison task of the PAN 2012 plagiarism detection competition[3] as part of the PAN series [3]. For the training phase, an evaluation service is provided where the participants upload their results and receive the performance score. For the final performance assessment, the participants submit their detection algorithms as executable software on either a Windows 7 or Linux based virtual machine. All submissions are automatically evaluated on the holdout test set with the reference implementation. This method is required since the organizers evaluate the detection approaches using real data that is subject to non-disclosure. The proposed method also allows the runtime of the submitted approaches to be recorded for the first time. The participants also have the possibility to opt-in for a public release of their plagiarism detection software as a service.

4. SUMMARY

In this paper, we proposed a web framework for IR experiments as a service motivated by low trends in sharing data and software in recent IR research. The fundamental design decisions concerning local instantiation, web dissemination, platform independence, result retrieval, and peer to peer collaboration address the specific needs of IR research. A reference implementation has been developed complying with these design rules that has been put to widespread use as part of the PAN competition series in 2012.

References

[1] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: Ad-hoc Retrieval Results since 1998. In CIKM '09, pages 601–610, Hong Kong, China, 2009.

[2] D. Lagun and E. Agichtein. ViewSer: Enabling Large-Scale Remote User Studies of Web Search Examination and Interaction. In SIGIR '11, pages 365–374, Beijing, China, 2011.

[3] M. Potthast. *Technologies for Reusing Text from the Web*. PhD Thesis, Bauhaus-Universität Weimar, Weimar, Germany, 2011.

[2] http://tira.webis.de
[3] http://pan.webis.de