Repeatability & Workability Evaluation of SIGMOD 2009*

I. Manolescu[1] S. Manegold[2] L. Afanasiev[1] J. Feng[3] G. Gou[4]
M. Hadjieleftheriou[5] S. Harizopoulos[6] P. Kalnis[7] K. Karanasos[1] D. Laurent[8]
M. Lupu[9] N. Onose[10] C. Ré[11] V. Sans[8] P. Senellart[12] T. Wu[13]
D. Shasha[14]

[1] INRIA Saclay–Île-de-France, France
[2] CWI, Netherlands
[3] University of Amsterdan, Netherlands
[4] Sun Yat-Sen University, China
[5] AT&T Shannon Labs, USA
[6] HP Labs, USA
[7] KAUST, Saudi Arabia
[8] Université de Cergy-Pontoise, France
[9] National U. of Singapore, Singapore
[10] University of California, San Diego, USA
[11] University of Washington, USA
[12] Télécom Paristech, France
[13] University of Illinois at Urbana-Champaign, USA
[14] Courant Institute, New York, USA

ABSTRACT

SIGMOD 2008 was the first database conference that offered to test submitters' programs against their data to verify the repeatability of the experiments published [1]. Given the quite positive experiences with and feedback about the SIGMOD 2008 repeatability initiative, SIGMOD 2009 modified and expanded the initiative with a workability assessment.

1. THE GOAL

On a voluntary basis, authors of accepted SIGMOD 2009 papers provided their code/binaries, experimental setups and data to be tested for:

repeatability of the experiments described in the accepted papers;

workability in the sense of running different/more experiments with different/more parameters than shown in the respective papers;

by a repeatability/workability committee (which we call the *RWC*), under the responsibility of the repeatability/workability editors-in-chief (which we call the *RWE*).

2. THE PEOPLE

The RWE are Ioana Manolescu and Stefan Manegold. The 2009 RWC consisted of the other authors of this paper, with the exception of D. Shasha.

*`http://homepages.cwi.nl/~manegold/SIGMOD-2009-RWE/`

3. THE PLAN

Several lessons learned from the first repeatability evaluation with SIGMOD 2008 [1] led us to improve and extend the process. The following paragraphs describe the details.

3.1 Accepted papers, only

The SIGMOD 2009 repeatability & workability evaluated accepted papers only. The primary reason for this change was to reduce the workload for the evaluation by avoiding evaluation of papers that would eventually not be accepted. A second reason was that authors had commented that they wouldn't mind the extra work of preparing their repeatability & workability submission in case their papers were accepted.

3.2 Adapted schedule

Focussing on accepted papers only required an adaptation of the general schedule for the repeatability & workability evaluation. After the SIGMOD 2009 program committee had announced the accepted papers, the contact authors of all accepted research papers were personally invited via email to prepare and submit their experiments including code, data sets and detailed instructions. This later start of the evaluation did not leave enough time to finish the evaluation before the camera ready deadline to allow the authors to mention the result of the evaluation in the final versions of their papers. Instead, the evaluation was completed just before the conference to give the authors the chance to mention the results in their presentations at SIGMOD 2009.

3.3 Refined submission method

In contrast to the push-based submission in 2008 via upload to a FTP server, the submission in 2009 was pull-based. Authors were asked to make their submissions available for download by the RWE. This helped to avoid problems with uploading large (tens of gigabytes) submissions to a single FTP server. The RWE then made the submissions available for download to the assigned reviewers.

3.4 Refined submission instructions

To give the reviewers some information to better plan their evaluation, the authors were asked to include in their

submission information about how long their experiments were expected to run. In addition, to facilitate the workability evaluation, the authors were asked to extend their repeatability instructions with suggestions as to how to extend their experiments beyond the contents of their paper. Possibilities ranged from explanations of how to use different data sets, query work loads, tuning- and/or configuration-parameters to compilation and installation instructions for alternative hardware-/software- environments.

3.5 Refined reviewing process

As last year, the assignment of papers to reviewers was mainly determined by the need to match the papers' hardware and software requirements with the reviewers resources. Of course, (potential) conflicts of interest were avoided. Unlike last year, each paper was assigned two reviewers. A *primary* reviewer to do the actual repeatability and workability evaluation, and a *secondary* reviewer as back-up and to double-check the primary reviewers report.

3.6 Author-reviewer-interaction

The 2008 experiences revealed that successful repeatability evaluation can be unnecessarily complicated, hindered or even prevented by minor problems with setting up and running the experiments due to minor unforeseen problems and/or missing details in the provided instructions. To solve this problem, the 2009 edition provided a web-base anonymous communication channel to allow interaction between authors and reviewers to resolve such minor problems as early as possible. All communication has been archived. With standard WIKI or BLOG software either not providing convenient and effective means for anonymous peer-to-peer communication or being considered an "overkill" for this task, a self-written PHP script was used to efficiently provide the basic functionality required.

4. THE PROCESS

After the announcement of the accepted papers, the contact authors of all 64 accepted research papers were invited by email to prepare and submit their contribution. By the (extended) deadline on April 22 2009, only 19 authors had reacted and provided their contribution. The remaining 45 authors chose to not reply at all. As opposed to last year, authors were not asked to provide in explanation why they could not submit their code, data and experiments for evaluation.

Each RWC member was assigned three papers, either two for primary review and one for secondary review, or one for primary and two for secondary review. Assigned reviewers were able to satisfy all software and hardware requirements, though sometimes at significant effort. For example, some reviewers needed to install extra software or even (re-)install complete machines. In one case, the reviewer's group even (re-)installed a 40-node Linux PC cluster to repeat a scaled-down version of experiments that were originally run on a 100-node cluster.

In nearly all cases, the anonymous web-base communication channel between authors and reviewers was intensively used to discuss and solve mostly minor problems — ranging from missing gnuplot files to not accurately enough specified versions of required software. Only in two cases, the discussion could not fully solve all problems, resulting in only a partial repeatability evaluation for those papers.

The reviewing process stretched over a complete two month period, with the final reviews being finished only the day before SIGMOD 2009 started. The long time was partly due to the extra hardware and software installation and configuration work required as mentioned above, partly due to delays in the author-reviewer communication to solve initial problems, and partly due to experiments that took several days or even weeks to run.

Although asked for, not all authors provided hints how to modify and/or extend their experiments for workability evaluation. In all but 5 cases, the reviewers managed to find their own ways to modify/extend the respective experiment to assess their workability. Also in cases with workability suggestions provided the reviewers volunteered to go beyond the authors' suggestions.

5. THE RESULTS

Overall, the results of the evaluation for the 19 submissions are rather positive:

- For 10 papers, the presented experiments could be fully repeated and workability was confirmed.
- For 1 paper, repeatability was fully confirmed and workability was mostly confirmed.
- For 4 papers, all original experiments were successfully repeated, but workability was not, due to missing or insufficient instructions on how to modify the original setup conveniently.
- For 1 paper, the experiments were mostly repeated, but workability could not be evaluated.
- For 1 paper, the repeatability evaluation was successful, but the workability evaluation failed.
- For 2 papers, major technical problems could not be solved within the 2 months reviewing period, preventing most or all of the repeatability and workability evaluation.

The authors were informed (just) before the conference about the results for their papers, and thus given the opportunity to mention the results during their presentation at SIGMOD 2009.

6. THE ASSESSMENT

With many of the lessons learned from the 2008 edition, the 2009 repeatability and workability went much smoother than the previous round. In particular focussing on accepted papers only (an idea suggested by Donald Kossmann), pull-based submission and the possibility for discussions between reviewers and authors to solve minor technical problems proved to be successful.

Though creating a higher workload for the reviewers, the newly introduced workability evaluation (if successful) gave even more credibility to the authors than a pure repeatability evaluation.

The unexpectedly low submission rate appears to be due to the fact that the authors were not aware of the SIGMOD 2009 repeatability & workability evaluation by the paper submission deadline. Due to several delays and issues, the SIGMOD 2009 repeatability & workability evaluation was not announced in any call for papers, nor mentioned

on the SIGMOD 2009 web site. Several personal communications with authors during the conference revealed that many authors were caught by surprise when invited to submit repeatability material for their accepted papers, or were simply not sure how "official" the evaluation was. In other words, there was probably insufficient publicity around the SIGMOD 2009 repeatability & workability evaluation.

While serving its primary purpose, the PHP script for the reviewer-author-communication has some room for improvements. Not being a standard tool, the "look-and-feel" was considered "unusual" and the automatic email notification of new postings did not always work reliably.

Given the diversity of the papers and their experiments, the reviewers were not given a strict format for their reviews, but rather allowed to freely determine the format, structure and content of their reviews themselves, to accommodate the process they followed as well as their findings and final verdict.

7. THE RECOMMENDATIONS

While overall considered a success and an improvement over the first edition in 2008, also the second edition in 2009 holds some lessons to be followed with future editions.

- The process of publicizing the RW process to the community needs to be improved, clearly mentioning the repeatability & workability evaluation on the SIGMOD web site, as well as announcing it in the call for papers.

- The author instructions need be improved further, in particular to ask more explicitly for workability instructions. More generally, collecting, improving and disseminating guidelines for the preparation of repeatable experiments requires more work in the community; tutorials such as [2] are a step in this direction.

- The review guidelines and format need to be improved and unified. Given the diversity of the experiments, this is not a trivial task, as any guidelines and/or format still need to leave sufficient room for all cases.

- The visibility of the RW evaluation results may be improved. In particular, it would be helpful to publish them on a public Web site, possibly together with the code that the authors may be willing to share. The SIGMOD PubZone server [3] is a promising tool for this purpose.

- The software support for author-reviewer discussions needs to be improved.

With respect to the last item above, we are currently considering the extension of the MyReview conference management tool [5] to accomodate the specific needs or our process. The main feature we need, and which is not yet supported by MyReview and other similar tools, is the possibility for reviewers and authors to exchange an unbound number of messages, over the whole period of reviewing (as opposed to one single exchange, at a specific point in the process, as currently used for conferences such as ACM SIGMOD, and supported by the Microsoft Research Conference Management Tool [4]).

8. REFERENCES

[1] I. Manolescu, L. Afanasiev, A. Arion, J.-P. Dittrich, S. Manegold, N. Polyzotis, K. Schnaitter, P. Senellart, S. Zoupanos, and D. Shasha. The repeatability experiment of SIGMOD 2008. *SIGMOD Record*, 37(1):39–45, Mar. 2008.

[2] I. Manolescu and S. Manegold. Performance Evaluation in Database Research: Principles and Experience. In *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, Cancún, Mexico, 2008. Tutorial slides are available from `http://www.icde2008.org/` or from the authors. A shortened version was presented also at the EDBT 2009 conference.

[3] PubZone: scientific publication discussion forum. `http://www.pubzone.org`.

[4] The Microsoft Research Conference Management Tool. `https://cmt.research.microsoft.com`.

[5] The MyReview Conference Management System. `http://myreview.lri.fr`.