Report on the Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys)

Alejandro Bellogín[1,3] Pablo Castells[3] Alan Said[1] Domonkos Tikk[2]
[1] Centrum Wiskunde & Informatica, The Netherlands
[2] Gravity R&D, Hungary
[3] Universidad Autónoma de Madrid, Spain

Abstract

Experiment replication and reproduction are key requirements for empirical research methodology, and an important open issue in the field of Recommender Systems. When an experiment is repeated by a different researcher and exactly the same result is obtained, we can say the experiment has been replicated. When the results are not exactly the same but the conclusions are compatible with the prior ones, we have a reproduction of the experiment. Reproducibility and replication involve recommendation algorithm implementations, experimental protocols, and evaluation metrics. While the problem of reproducibility and replication has been recognized in the Recommender Systems community, the need for a clear solution remains largely unmet, which motivates the main questions addressed in the present workshop.

1 Introduction

The *Reproducibility and Replication in Recommender Systems Evaluation* (RepSys) Workshop[1] [3] was held on the 12th October 2013 in conjuction with the 7th ACM Recommender Systems conference (RecSys), at Hong Kong, China. We report here the main activities and discussions that took place.

The empirical evaluation of Recommender Systems (RS) is acknowledged to be an open problem in the field, with open issues yet to be addressed [13]. Many experimental approaches and metrics have been developed along the years, which the community is well acquainted with, but key aspects and details in the design and application of available methodologies are open to configuration and interpretation, where even apparently subtle details may create a considerable difference. This results in a significant divergence in experimental practice, hindering the comparison and proper assessment of contributions and advances to the field.

In this context, the replication and reproduction of experiments is one of the desirable requirements for experimental research still to be met in the field. We say an experiment is replicated when it is repeated by a different researcher and exactly the same result is

[1] `http://repsys.project.cwi.nl`

obtained. When the results are not exactly the same but the conclusions are compatible with the prior ones, we have a reproduction of the experiment.

The topic of reproducibility is an obvious concern at this moment in several fields, such as Information Retrieval and Human-Computer Interaction (e.g. RepliCHI panel in 2012 and workshop in 2013 [16]). Adjacent to this issue are the workshops dealing with software engineering in recommendation (RSSE series [7] and book [12]) and open source software, again in Information Retrieval (Open Source Information Retrieval workshops at SIGIR 2006 and 2012 [15,17]) and Machine Learning (Machine Learning Open Source Software workshops at NIPS 2006 and 2008 [14]).

The discussion and definition of the basic elements of the experimental conditions (and their requirements) are critical to support continuous innovation in any discipline. The offline evaluation of recommender systems requires an implementation of the algorithm or technique to be evaluated, a set of quality measures for comparative evaluation, and an experimental protocol establishing how to handle the data and compute metrics in detail. Online evaluation similarly requires an algorithm implementation and a population of users to survey (by means of an A/B test, for instance). Here again, perhaps even more importantly than in offline evaluation, an experimental protocol needs to be established and adhered to. As a paradigmatic example, the Information Retrieval field, adjacent to RS, has established a solid and successful development in such terms with the TREC conference[2] and shared public scripts (e.g. treceval) to evaluate systems on the tasks proposed every year in that venue.

Even when a set of publicly available resources (data and algorithm implementations) exists in the RS community, very often research studies do not report comparable results for the same methods *under the same conditions*. This is due to the high number of experimental design options and parameters in recommender system evaluation, and the huge impact of the experiment configuration on the outcomes.

In order to seek reproducibility and replication several strategies can be considered, such as source code sharing, standardization of agreed evaluation metrics and protocols, or releasing public experimental design software, all of which have difficulties of their own. Furthermore, for online evaluation, an extensive analysis of the population of test users should be provided. While the problem of reproducibility and replication has been recognized in the community, the need for a solution remains largely unmet. This, together with the need for further discussion, methodological standardization for both reproducibility and replication motivates the issues addressed in the workshop. The following sections report the topics discussed in the invited keynote, accepted papers, and an interactive panel where experienced members of the academic and industrial communities discussed the challenges and problems of reproducibility and replication in recommender systems.

2 Keynote

The workshop opened with a keynote by Mark Levy (Mendeley) entitled *Offline evaluation of recommender systems: all pain and no gain?* [10]. The talk started by considering the differences between a good and a bad recommendation; the main problem being that it is difficult to measure directly and decide why a recommendation was good or bad. For instance, a good recommendation can be defined as one that increases the usefulness of your product

[2] `http://trec.nist.gov`

in the long run but this is very hard to measure; on the other hand, it is quite obvious for a user to know when she receives a bad recommendation, although it is very difficult to infer why (not relevant for the user, too obscure, too familiar, the user already has it or already knows she does not like it, badly explained, etc.). Related with this is the cost of getting a bad recommendation, which will mainly depend on the product and the users of the system.

Mark presented in the keynote two relevant hypothesis when evaluating real recommendations in an offline context: good offline metrics express product goals, and most (really) bad recommendations can be caught by business logic. However, there are other issues which need to be considered, specifically, that real business goals tend to be concerned about the long-term user behavior (e.g., Netflix), although short-term surrogates are usually exploited instead; besides, only partial user behavior is visible. These constraints, it should be noted, are the same when training data are collected. In this context, the *least bad solution* would include the analysis of historical logs, deciding which events indicate success, emphasizing precision at higher cutoffs, and using recall to discriminate once precision reaches a plateau.

Additionally, there should be an increased effort in making metrics meaningful. Evaluation should be realistic (leaving the ivory tower) and, at the same time, provide test setups and reproducible baselines. For this, we need an honest measure of the preference (predicted items may not be correct just because they were consumed), we should capture the value of the recommendation (it is hard to say if a recommender that is useful in the short term may be just too obvious) and not neglect (contextual) side-data. As a conclusion, public data alone may not be enough to guarantee reproducibility or the fair comparison of methods. Importantly, preparation and evaluation code should be released as well.

Finally, he discussed the motivation of offline evaluation. Two examples (movie recommendation based on error minimization and reducing playlist skips) were presented as a proof of concept where poor offline evaluation can lead to years of misdirected research. Contests like the Yahoo! Music KDD [6], the million song challenge [4], or the Yelp RecSys challenge [5] can help by enforcing consistent evaluation, however privacy concerns usually result in obfuscated – and not so useful – data. Because of this, the focus should be on building test frameworks that ensure clear offline goals, while at the same time, help to derive efficient online tests, by cutting down the huge parameter space.

Mark also mentioned his own experience when building a small framework around a new algorithm[3]. He pointed out that usable frameworks are hard to write, mainly because there is a tradeoff between clarity and scalability. Additionally, he proposed a wishlist for a framework aimed to produce reproducible evaluation: enable integration with other recommender implementations, handle data formats and pre-processing, handle splitting, cross-validation, side datasets, save everything to file, generate meaningful metrics, be well documented and easy to use. Among the current offerings (GraphChi/GraphLab, Mahout, Lenskit, MyMediaLite) none of them meet all these requirements.

3 Paper presentations

Andrej Košir from University of Ljubljana talked about "How to improve the statistical power of the 10-fold cross validation scheme in Recommender Systems" [9]. The authors propose a procedure to detect which contextual variables are relevant when a cross validation setting is used. The actual problem they observe has to do with reproducibility, not replicability;

[3] `https://github.com/mendeley/mrec`

the problem appears when performing statistical tests on each of the separate folds and an additional test on the complete results, and what a researcher should do when there are significant differences in one of the cases but not in the other.

Stefan Langer, from Docear, presented two papers about research paper recommendation. The first one was a survey entitled "Research paper recommender system evaluation: a quantitative literature survey" [2], where the authors reviewed more than 170 research papers about academic literature recommendation. They found that no consensus exists at the moment about how to evaluate and compare this type of systems – in particular, 21% of the approaches were not evaluated at all or used uncommon methods.

The second paper, entitled "A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation" [1] provided a comparison between offline and online evaluation results, again, in the context of research paper recommendation. The authors note that contradicting results are found when comparing performance measurements using clickthrough rate, mean average precision, and precision at 10. Besides, offline evaluation does not (always) show enough predictive power with respect to online evaluation.

The last presentation was given by Joseph Konstan from the University of Minnesota, "Toward identification and adoption of best practices in algorithmic recommender systems research" [8]. In this paper the authors discuss that *doing research right is hard*. They argue that what makes research believable is to have internal (correct research, proper sampling, not confounding factors) and external (realistic scenario) validity. Joseph suggested the use of checklists in recommender systems research – by adapting Atul Gawande's idea on *The checklist manifesto*. In this way, a list of community-generated best practices would be presented to authors and reviewers to think about the quality of the paper, to be considered as guidelines, not as requirements.

4 Panel Session

As the last part of the workshop, we invited various experts to join our panel session to discuss reproducibility and replication in recommendation. We were happy to welcome Paolo Cremonesi (Politecnico di Milano, Moviri), our keynote speaker Mark Levy, and the three presenters, Joseph Konstan, Andrej Košir, and Stefan Langer.

The first question for the panel was if reproducibility and replication are important in recommender systems. It was discussed that they are, indeed, but not very important; what is really important is theory building. The issue of which parts of papers should be replicated was also raised: in most cases the critical result, but it is also important to try different scenarios and user setups, otherwise it is not possible to connect the results with the real world.

Another addressed issue was the reproducibility of online experiments. They are not reproducible and it would be very hard to create an open system for that, in principle because there is not a generic solution to do online A/B testing and because by running such a test we can only guarantee that some statistical features of the user groups are similar or different. In any case, there has to be some connection between online and offline evaluation. In particular, offline settings should fit into online evaluation, that is, they should be highly related to the real world. A possible reason for the small correlation between offline and online results is probably not in the metric itself, but all the information we are missing from

the big picture – assumptions such as missing ratings not at random [11] may cause these metrics not to be measured in a reliable way, and thus we may need to rethink them.

A recurrent topic was how to foster reproducibility. As in general computer science, we should publish the source code, especially if we use a public dataset. However, overfitting may cause authors not to share their code. It was also mentioned that the industry should also embrace reproducibility, in part because the algorithm is not the only secret they have to protect, there are several other aspects. In this context, industry should, at least, try their algorithms also on public datasets.

Another side of this problem is that we do not want to raise the bar of publishing a paper in the conference too much. There should be enough value in the paper to integrate other changes or ideas. Usually a paper claiming that the authors could not replicate another paper does not get published; however, this (negative) result is important, since it may – eventually – show that there is an issue with that specific paper (in the algorithm being presented, how the evaluation is performed, or how the data were processed). As a final note of the workshop, the panelists suggested that we should distinguish negative results (still not easy to publish / find the value within the community) from refutations (a result that contradicts – or not – a published paper), probably as a separate track in the main conference.

5 Conclusions and Future Directions

The workshop featured an excellent keynote and a range of interesting papers from both academia and industry. Several aspects of evaluation in recommender systems were presented and discussed, focused on the main topics of the workshop: reproducibility and replication. The RepSys workshop was successful in bringing people from the research community and industry together, and in discussing and addressing some important issues in the area, emphasizing the importance of (proper) evaluation and the need for some guidelines to produce better (useful, interesting, significant) research results.

6 Acknowledgements

We would like to thank the organizers of ACM RecSys for providing a venue for this workshop. Furthermore, we acknowledge the efforts of the members of the program committee, including: Xavier Amatriain (Netflix, USA), Linas Baltrunas (Telefonica Research, Spain), Marcel Blattner (University of Applied Sciences, Switzerland), Iván Cantador (Universidad Autónoma de Madrid, Spain), Ed Chi (Google Inc., USA), Arjen de Vries (Centrum Wiskunde & Informatica, The Netherlands), Juan Manuel Fern'andez (Universidad de Granada, Spain), Zeno Gantner (Nokia, Germany), Pankaj Gupta (Twitter, USA), Andreas Hotho (University of Würburg, Germany), Juan Huete (Universidad de Granada, Spain), Kris Jack (Mendeley, UK), Dietmar Jannach (University of Dortmund, Germany), Jaap Kamps (University of Amsterdam, The Netherlands), Alexandros Karatzoglou (Telefonica Research, Spain), Bart Knijnenburg (University of California, Irvine, USA), Ido Guy (IBM Haifa Research Lab, Israel), Jérome Picault (Bell Labs, Alcatel-Lucent, France), Till Plumbaum (TU Berlin, Germany), Daniele Quercia (Yahoo!, Spain), Filip Radlinski (Microsoft, Canada), Yue Shi (TU-Delft, The Netherlands), Fabrizio Silvestri (Consiglio

Nazionale delle Ricerche, Italy), Harald Steck (Netflix, USA), David Vallet (NICTA, Australia), Jun Wang (University College London, UK), Xiaoxue Zhao (University College London, UK).

Special thanks are due to the paper authors, the invited speaker, and all the participants for a lively workshop.

We would like to thank the ERCIM "Alain Bensoussan" Fellowship Programme, funded by European Comission FP7 grant agreement no.246016, the Centrum Wiskunde & Informatica, and Universidad Autónoma de Madrid for their financial contributions.

References

[1] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13, pages 7–14, New York, NY, USA, 2013. ACM.

[2] Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, Corinna Breitinger, and Andreas Nürnberger. Research paper recommender system evaluation: A quantitative literature survey. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13, pages 15–22, New York, NY, USA, 2013. ACM.

[3] Alejandro Bellogín, Pablo Castells, Alan Said, and Domonkos Tikk, editors. *RepSys '13: Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, New York, NY, USA, 2013. ACM.

[4] Thierry Bertin-Mahieux, Daniel P. W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *ISMIR*, pages 591–596, 2011.

[5] Jim Blomo, Martin Ester, and Marty Field. Recsys challenge 2013. In *RecSys*, pages 489–490, 2013.

[6] Gideon Dror, Noam Koenigstein, Yehuda Koren, and Markus Weimer. The yahoo! music dataset and kdd-cup '11. In *KDD Cup*, pages 8–18, 2012.

[7] Reid Holmes, Martin P. Robillard, Robert J. Walker, and Thomas Zimmermann. Rsse 2010: Second international workshop on recommendation systems for software engineering. In *ICSE (2)*, pages 455–456, 2010.

[8] Joseph A. Konstan and Gediminas Adomavicius. Toward identification and adoption of best practices in algorithmic recommender systems research. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13, pages 23–28, New York, NY, USA, 2013. ACM.

[9] Andrej Košir, Ante Odić, and Marko Tkalčič. How to improve the statistical power of the 10-fold cross validation scheme in recommender systems. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13, pages 3–6, New York, NY, USA, 2013. ACM.

[10] Mark Levy. Offline evaluation of recommender systems: All pain and no gain? In *Proceedings of the International Workshop on Reproducibility and Replication in Rec-*

ommender Systems Evaluation, RepSys '13, pages 1–1, New York, NY, USA, 2013. ACM.

[11] Benjamin M. Marlin, Richard S. Zemel, Sam T. Roweis, and Malcolm Slaney. Collaborative filtering and the missing at random assumption. In Ronald Parr and Linda C. van der Gaag, editors, *UAI*, pages 267–275. AUAI Press, 2007.

[12] Martin P. Robillard, Walid Maalej, Robert J. Walker, and Thomas Zimmermann, editors. *Recommendation Systems in Software Engineering*. Springer, 2014.

[13] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender Systems Handbook*, pages 257–297. 2011.

[14] Sören Sonnenburg, Mikio L. Braun, Cheng Soon Ong, Samy Bengio, Léon Bottou, Geoffrey Holmes, Yann LeCun, Klaus-Robert Müller, Fernando Pereira, Carl Edward Rasmussen, Gunnar Rätsch, Bernhard Schölkopf, Alexander J. Smola, Pascal Vincent, Jason Weston, and Robert C. Williamson. The need for open source software in machine learning. *Journal of Machine Learning Research*, 8:2443–2466, 2007.

[15] Andrew Trotman, Charles L. A. Clarke, Iadh Ounis, Shane Culpepper, Marc-Allen Cartright, and Shlomo Geva. Open source information retrieval: a report on the SIGIR 2012 workshop. *SIGIR Forum*, 46(2):95–101, 2012.

[16] Max L. Wilson, Ed H. Chi, David Coyle, and Paul Resnick, editors. *Proceedings of the CHI 2013 Workshop on the Replication of HCI Research, Paris, France, April 27-28, 2013*, volume 976 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[17] Wai Gen Yee, Michel Beigbeder, and Wray L. Buntine. SIGIR06 workshop report: Open source information retrieval systems (OSIR06). *SIGIR Forum*, 40(2):61–65, 2006.