Repositorio Institucional de la Universidad Autónoma de Madrid

https://repositorio.uam.es

Esta es la **versión de autor** de la comunicación de congreso publicada en:
This is an **author produced version** of a paper published in:

RecSys '13: Proceedings of the 7th ACM conference on Recommender systems, ACM, 2013. 485-486

DOI: http://dx.doi.org/10.1145/2507157.2508006

Copyright: © 2013 ACM

El acceso a la versión del editor puede requerir la suscripción del recurso
Access to the published version may require subscription

Workshop on Reproducibility and Replication in Recommender Systems Evaluation – RepSys

Alejandro Bellogín*, Pablo Castells†, Alan Said*, Domonkos Tikk‡
alejandro.bellogin@cwi.nl, pablo.castells@uam.es, alan.said@cwi.nl,
domonkos.tikk@gravityrd.com

*Centrum Wiskunde & Informatica, Science Park 123, 1098 XG Amsterdam, The Netherlands
†Universidad Autónoma de Madrid, Francisco Tomás y Valiente 11, 28049 Madrid, Spain
‡Gravity R&D, Expo tér 5-7, 1101 Budapest, Hungary

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 present workshop.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: information filtering, relevance feedback, retrieval models, search process, selection process.

General Terms

Algorithms, Design, Experimentation, Measurement, Performance

Keywords

Evaluation, replicability, reproducibility, experimental design, experimental methodology

1. INTRODUCTION

The empirical evaluation of Recommender Systems (RS) is acknowledged to be an open problem in the field, with open issues yet to be addressed [2]. 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 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 (RepliCHI panel in 2012 and workshop in 2013 [5]). Adjacent to this issue are the workshops dealing with software engineering in recommendation (RSSE series [1] and prospective book for 2013[1]) and open source software, again in Information Retrieval (Open Source Information Retrieval workshops at SIGIR 2006 and 2012 [4, 6]) and Machine Learning (Machine Learning Open Source Software workshops at NIPS 2006 and 2008 [3]).

The discussion and definition of the basic elements of the experimental conditions (and their requirements) is critical to support continuous innovations 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, is a successful development on this ground, with the TREC conferences[2] and a common tool (treceval) to evaluate any of the tasks proposed 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 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
RecSys'13, October 12–16, 2013, Hong Kong, China.
Copyright is held by the owner/author(s).
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

[1] https://sites.google.com/site/rsseresearch/rsse-book
[2] http://trec.nist.gov

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 in both reproducibility as well as replication motivates the present workshop.

2. SCOPE AND GOALS

The workshop gathered researchers and practitioners interested in defining clear guidelines for their experimental needs to allow fair comparisons to related work. The workshop provided an informal setting for exchanging and discussing ideas, sharing experiences and viewpoints. We aimed to identify and better understand the current gaps in the implementation of recommender system evaluation methodologies, help lay directions for progress in addressing them, and foster the consolidation and convergence of experimental methods and practice. The workshop sought to identify the main challenges related to reproduction and replication of prior research, along with an exploration of possible directions to overcome these limitations.

Specific questions raised and addressed at the workshop includes the following:

- How important is the reproducibility and replication of experiments for the RS community?
- What are the challenges for replication of evaluation in the RS field? How could we facilitate easier and more accurate comparison with prior work?
- How can methods and metrics be more clearly and/or formally defined within specific tasks and contexts for which a recommender application is deployed?
- What parts –if any– of an online experiment could be reproducible and how?
- How should the academic evaluation methodologies be described to improve their relevance, usefulness, and replicability for industrial settings?
- What type of public resources (data sets, benchmarks) should be available, and how can they be built? Is it possible to have a generic framework for the evaluation (and replication) of recommender systems?
- To what extent is it possible to reuse experimental methodologies across domains and/or businesses?
- How do we envision the evaluation of recommender systems in the future and how does this affect the replicability of said systems?

3. COVERED TOPICS

The accepted papers and the discussions held at the workshop addressed –among others– the following topics:

- Limitations and challenges of experimental reproducibility and replication
- Reproducible experimental design
- Replicability of algorithms
- Standardization of metrics: definition and computation protocols
- Evaluation software: frameworks, utilities, services
- Reproducibility in user-centric studies
- Datasets and benchmarks
- Recommender software reuse
- Replication of already published work
- Reproducibility within and across domains and organizations
- Reproduction and replication guidelines

4. OVERVIEW

The workshop took place on October 12th, 2013. It opened with a keynote talk by Mark Levy (Mendeley), followed by the presentation of accepted papers and open discussions, where an interactive panel with prominent members of the academic and industrial communities discussed the challenges and problems of reproducibility and replication in recommender systems. The accepted papers and a summary of discussions are available in the workshop proceedings published in the ACM International Conference Proceedings Series (ICPS), which can be reached from the workshop website at http://repsys.project.cwi.nl.

5. ACKNOWLEDGMENTS

This workshop was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme, funded by European Comission FP7 grant agreement no.246016.

6. REFERENCES

[1] HOLMES, R., ROBILLARD, M. P., WALKER, R. J., AND ZIMMERMANN, T. Rsse 2010: Second international workshop on recommendation systems for software engineering. In *ICSE (2)* (2010), J. Kramer, J. Bishop, P. T. Devanbu, and S. Uchitel, Eds., ACM, pp. 455–456.

[2] SHANI, G., AND GUNAWARDANA, A. Evaluating recommendation systems. In *Recommender Systems Handbook*. 2011, pp. 257–297.

[3] SONNENBURG, S., BRAUN, M. L., ONG, C. S., BENGIO, S., BOTTOU, L., HOLMES, G., LECUN, Y., MÜLLER, K.-R., PEREIRA, F., RASMUSSEN, C. E., RÄTSCH, G., SCHÖLKOPF, B., SMOLA, A. J., VINCENT, P., WESTON, J., AND WILLIAMSON, R. C. The need for open source software in machine learning. *Journal of Machine Learning Research 8* (2007), 2443–2466.

[4] TROTMAN, A., CLARKE, C. L. A., OUNIS, I., CULPEPPER, S., CARTRIGHT, M.-A., AND GEVA, S. Open source information retrieval: a report on the SIGIR 2012 workshop. *SIGIR Forum 46*, 2 (2012), 95–101.

[5] WILSON, M. L., CHI, E. H., COYLE, D., AND RESNICK, P., Eds. *Proceedings of the CHI 2013 Workshop on the Replication of HCI Research, Paris, France, April 27-28, 2013* (2013), vol. 976 of *CEUR Workshop Proceedings*, CEUR-WS.org.

[6] YEE, W. G., BEIGBEDER, M., AND BUNTINE, W. L. SIGIR06 workshop report: Open source information retrieval systems (OSIR06). *SIGIR Forum 40*, 2 (2006), 61–65.