Replicable Evaluation of Recommender Systems

Alan Said
Recorded Future
Sweden
alansaid@acm.org

Alejandro Bellogín
Universidad Autónoma de Madrid
Spain
alejandro.bellogin@uam.es

ABSTRACT

Recommender systems research is by and large based on comparisons of recommendation algorithms' predictive accuracies: the better the evaluation metrics (higher accuracy scores or lower predictive errors), the better the recommendation algorithm. Comparing the evaluation results of two recommendation approaches is however a difficult process as there are very many factors to be considered in the implementation of an algorithm, its evaluation, and how datasets are processed and prepared.

This tutorial shows how to present evaluation results in a clear and concise manner, while ensuring that the results are comparable, replicable and unbiased. These insights are not limited to recommender systems research alone, but are also valid for experiments with other types of personalized interactions and contextual information access.

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 Recommender System community strives towards improving the quality of recommendation algorithms, in order to do so, it is imperative that comparisons across recommendation approaches can be performed in an accurate and unbiased fashion. Assuming the assumption "the higher the evaluation scores, the better the recommender algorithm", which is usually taken for granted, it is important for both researchers and practitioners that their evaluations are fair. However, it is difficult to compare the results from a given evaluation of a recommender system, mainly because the very many alternatives that exist in designing and implementing an evaluation strategy. At the ACM RecSys conferences, every year there usually are several papers on evaluation, additionally there have been a number of workshops (UCERSTI [5, 8], RUE [2], RepSys [4], REDD [1]) and tutorials on related topics. However, there has been little focus on the reproducibility and replicability of the evaluation and results themselves; hence, this tutorial aims to provide a broader view that integrates also this aspect into a general evaluation methodology.

A related tutorial was given at ACM Hypertext 2014 [3].

RecSys'15, September 16–20, 2015, Vienna, Austria.
ACM 978-1-4503-3692-5/15/09.
DOI: http://dx.doi.org/10.1145/2792838.2792841 .

2. TUTORIAL DESCRIPTION

This tutorial aims to give an introduction to clear and concise reporting of evaluation methodologies and results, while at the same time ensuring that the results of the evaluation are comparable, replicable and unbiased to allow for fair comparisons with related work. The tutorial defines and presents evaluation metrics, methodologies, and experimental configurations used in the recommender systems literature. Using these definitions, we present specific guidelines towards reporting experimental results in the recommender systems area. As a particular focus of interest, in the tutorial we address the commons datasets and benchmarking frameworks available, and how they can be applied in future publications in the recommender systems field in order to overcome limitations related to the lack of reproduction and reproducibility of the experiments and results.

2.1 Tutorial Structure

Introduction. This part of the tutorial focuses on the basics of recommendation and evaluation: core recommendation concepts, definitions of metrics and methodologies.

Evaluation. This section provides the necessary setting to understand how recommender systems are evaluated. A brief introduction to the basic evaluation concepts (metrics, data splits, etc.) allows participants on all levels to understand the basic setting. Following this, more specific concepts related to evaluation will be presented, e.g. data splitting criteria, biases that can arise from incorrectly configured algorithms, and calculations of metrics. We also discuss advanced evaluation concepts, such as subjective evaluation criteria (novelty, diversity) as well as methods used in in situ evaluation, e.g. A/B testing, significance testing, etc.

Replication. This section focuses on replication itself, i.e. how to best plan, perform, and report evaluation results in order

to allow for others to grasp the objective quality of an experiment without necessarily having to reproduce it themselves. We seek to present in a clear way specific guidelines towards reporting experimental results. Particular focus is put on common datasets and frameworks available, and show how they can be put to use in research publications in order to overcome limitations related to the lack of reproduction and reproducibility of the experiments.

Replication by example. This is an interactive session which presents results where several of the discussed configurations are tested with real data. The audience is invited to a discussion on expected results vs. obtained outcomes. The feedback obtained during the discussion is used to improve and augment the aforementioned experiments to other recommendation-related areas, such as contextual search or personalized mobile services. For this session, code examples are available on GitHub[1]. These code examples show the necessary steps in order to make the evaluation of recommender systems replicable and build on the recommender system evaluation toolkit RiVal [6, 7][2]. The demonstration combines common recommendation frameworks with RiVal in order to present an evaluation comparable across recommendation framework.

Conclusions and wrap-up. This session concludes the tutorial and iterates the most important factors to consider while planning and performing evaluation, not only for the sake of reproducibility by others, but also for the sake of correct and objective comparison within the same recommendation setting.

Q&A. This session gives the participants the opportunity to ask questions on the topics presented.

2.2 Intended Audience

The tutorial is designed to be useful for researchers, students, and practitioners in the Recommender Systems and personalization communities, and in related areas such as Information Retrieval, Data Mining, Machine Learning and Human-Computer Interaction, working in different application domains, and concerned with implementation, reproduction, evaluation, research and practice.

3. PRESENTERS

Alan Said is a Machine Learning engineer at Recorded Future in Gothenburg, Sweden. Prior to this he was a postdoctoral researcher at TU Delft, and Marie Curie (ERCIM ABCDE) Fellow at Centrum Wiskunde & Informatica in Amsterdam, The Netherlands. He obtained a doctorate at Technische Universität Berlin. His research interests include evaluation, benchmarking, user modeling and different aspects of recommender systems. He has served as PC member of international conferences and workshops (e.g. RecSys, UMAP, HT, IIiX, ECIR, ECMLPKDD, IUI) and as reviewer for journals (e.g. UMUAI, TIST, TKDD, TWEB). In 2010, 2011, 2012 and 2014 he co-organized the CAMRa and RecSys Challenge benchmarking challenges at ACM RecSys. At the 2012 ACM RecSys conference, Alan co-presented a tutorial on *Best Practices in Recommender System Challenges*. The tutorial outlined the necessary steps to be taken in order to achieve comparable results in the context of a competition or a challenge.

Alejandro Bellogín is a Lecturer at the Autónoma University of Madrid. Previously, he was an ERCIM Post-doctoral fellow at Centrum Wiskunde & Informatica in The Netherlands. His research is focused on recommender systems, in particular, adaptations from the information retrieval area, such as performance prediction techniques, evaluation methodologies, and probabilistic models. He has authored papers on national and international conferences, journals, and workshops in the aforementioned areas. He has co-organized a workshop on the topic of reproducibility in evaluation at the ACM RecSys conference. He has served as PC member of international conferences and workshops (such as CIKM, ECIR, and RecSys) and as reviewer for journals (e.g., IRJ, IPM, INS, TIST). At the 2014 ACM Hypertext conference, Alejandro Bellogín gave a tutorial on *Evaluating Recommender Systems - Ensuring Replicability of Evaluation*. The tutorial was tailored to a broad audience from other fields than recommendation.

[1]http://github.com/recommenders/evaltutorial

[2]http://rival.recommenders.net

4. ACKNOWLEDGMENTS

Supported in part by the Ministerio de Educación y Ciencia (TIN2013-47090-C3-2).

5. REFERENCES

[1] ADAMOPOULOS, P., BELLOGÍN, A., CASTELLS, P., CREMONESI, P., AND STECK, H. REDD 2014 - international workshop on recommender systems evaluation: dimensions and design. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014* (2014), pp. 393–394.

[2] AMATRIAIN, X., CASTELLS, P., DE VRIES, A. P., AND POSSE, C. Workshop on recommendation utility evaluation: beyond RMSE - RUE 2012. In *Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012* (2012), pp. 351–352.

[3] BELLOGÍN, A. Evaluating recommender systems:ensuring replicability of evaluation. http://ir.ii.uam.es/ alejandro/2014/ht.pdf, 2014.

[4] BELLOGÍN, A., CASTELLS, P., SAID, A., AND TIKK, D. Workshop on reproducibility and replication in recommender systems evaluation: Repsys. In *Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013* (2013), pp. 485–486.

[5] KNIJNENBURG, B. P., SCHMIDT-THIEME, L., AND BOLLEN, D. G. F. M. Workshop on user-centric evaluation of recommender systems and their interfaces. In *Proceedings of the 2010 ACM Conference on Recommender Systems, RecSys 2010, Barcelona, Spain, September 26-30, 2010* (2010), pp. 383–384.

[6] SAID, A., AND BELLOGÍN, A. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014* (2014), pp. 129–136.

[7] SAID, A., AND BELLOGÍN, A. Rival: a toolkit to foster reproducibility in recommender system evaluation. In *Eighth ACM Conference on Recommender Systems, RecSys '14, Foster City, Silicon Valley, CA, USA - October 06 - 10, 2014* (2014), pp. 371–372.

[8] WILLEMSEN, M. C., BOLLEN, D. G. F. M., AND EKSTRAND, M. D. UCERSTI 2: second workshop on user-centric evaluation of recommender systems and their interfaces. In *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011* (2011), pp. 395–396.