REDD 2014 – International Workshop on Recommender Systems Evaluation: Dimensions and Design

Panagiotis Adamopoulos
New York University
44 West Fourth Street
New York, NY 10012, USA
padamopo@stern.nyu.edu

Alejandro Bellogín,
Pablo Castells
Univ. Autónoma de Madrid
Fco. Tomás y Valiente 11
Madrid, 28049 Spain
alejandro.bellogin@uam.es
pablo.castells@uam.es

Paolo Cremonesi
Politecnico di Milano
Via Ponzio 34/5
Milan, Italy
paolo.cremonesi@polimi.it

Harald Steck
Netflix, Inc.
100 Winchester Circle
Los Gatos, CA 95032, USA
hsteck@netflix.com

ABSTRACT
Evaluation is a cardinal issue in recommender systems; as in any technical discipline, it highlights to a large extent the problems that need to be solved by the field and, hence, leads the way for algorithmic research and development in the community. Yet, in the field of recommender systems, there still exists considerable disparity in evaluation methods, metrics and experimental designs, as well as a significant mismatch between evaluation methods in the lab and what constitutes an effective recommendation for real users and businesses. Even after the relevant quality dimensions have been defined, a clear evaluation protocol should be specified in detail and agreed upon, allowing for the comparison of results and experiments conducted by different authors. This would enable any contribution to the same problem to be incremental and add up on top of previous work, rather than grow sideways. The REDD 2014 workshop seeks to provide an informal forum to tackle such issues and to move towards better understood and shared evaluation methodologies, allowing one to leverage the efforts and the workforce of the academic community towards meaningful and relevant directions in real-world developments.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: *Information filtering*

General Terms
Algorithms, Measurement, Performance, Experimentation, Standardization, Theory

Keywords
Utility, evaluation, methodology, metrics, recommender systems

1. INTRODUCTION
Thoughtful evaluation of recommender systems is a key challenge, as it guides algorithmic research and development [3,9]. In the community there is, however, considerable disparity in evaluation methods, metrics and experimental designs, as well as a significant mismatch between evaluation methods in the lab and what constitutes an effective recommendation for real users and businesses [5,6,9,11,12,13,15].

On the one hand, REDD 2014 places a specific focus on the identification and measurement of different recommendation quality dimensions that go beyond the monolithic concept of simply matching user preferences. Novelty and diversity, for instance, have been recognized as key perspectives of the utility of recommendations for users in real-world scenarios, with a direct positive effect on business performance [1,2,4,7,8,10,14]. Considering the business perspective, performance metrics related to sales, revenues, and user engagement along the recommendation funnel should also be used. Additionally, from an engineering point of view, aspects such as efficiency, scalability, robustness and user interface design are typically major concerns; often prioritized over the effectiveness of the internal algorithms at the core of the system.

On the other hand, once a relevant quality dimension has been defined, a clear evaluation protocol should be specified in detail. This is essential for reproducibility of experiments. Moreover, it enables different authors to build on top of other researchers' previous works. Even when measuring recommendation accuracy, researchers and practitioners are still often faced with experimental design questions for which there are not always precise and consensual answers. Therefore, there remains room for further methodological development and convergence, which motivates this workshop.

2. SCOPE AND GOALS
REDD 2014 gathered researchers and practitioners interested in better understanding the unmet needs in the field in terms of evaluation methodologies and experimental practices. The workshop provided an informal setting for exchanging and discussing ideas as well as sharing experiences and viewpoints. REDD sought to identify and better understand the current gaps in recommender system evaluation methodologies, help lay directions for progress in addressing them, and foster the consolidation and convergence of experimental methods and best practices.

Specific questions raised and addressed by the workshop include, among others, the following:

- What are the unmet needs and challenges for evaluation in the Recommender Systems field? Where do we stand? What changes would we like to see? How could we speed up progress?

RecSys'14, October 6–10, 2014, Foster City, Silicon Valley, CA, USA.
ACM 978-1-4503-2668-1/14/10.
http://dx.doi.org/10.1145/2645710.2645780

- What relevant recommendation utility and quality dimensions should be considered? How can they be captured and measured? How should evaluation methods be designed to effectively evaluate such dimensions?
- How can metrics become more clearly and/or formally related to the task, contexts, and goals for which a recommender application is deployed?
- How should IR metrics be applied to recommendation tasks? What aspects require adjustment or further clarification? What further methodologies should we draw from other disciplines (e.g., HCI, Machine Learning, etc.)?
- What biases and noise should experimental design typically watch for?
- Can we predict the success of a recommendation algorithm with offline experiments? What offline metrics correlate better and under which conditions?
- What are the outreach and limitations of offline evaluation? How can online and offline experiments complement each other?
- What type of public datasets and benchmarks would we want to have available, and how can they be built?
- How can the recommendation effect be traced on business outcomes?
- How should the academic evaluation methodologies improve their relevance and usefulness for industrial settings?
- How can we promote reproducibility of recommender systems methods?
- How do we envision the evaluation of recommender systems in the future?

3. COVERED TOPICS

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

- Evaluation methodology
- Experimental design
- Open evaluation platforms and infrastructures
- Recommendation quality dimensions: accuracy, novelty, diversity, unexpectedness, serendipity, coverage, risk, robustness, usability, explanations, persuasiveness, etc.
- Evaluating for efficiency and scalability
- Definition and assessment of evaluation metrics
- Matching metrics to tasks, needs, and goals
- Business-oriented evaluation
- Offline and online evaluation
- Datasets and benchmarks

The workshop opened with a keynote talk, followed by the presentation of accepted papers and open discussions. The accepted papers and a summary of discussions are available in the workshop proceedings, which can be reached from the workshop website at http://ir.ii.uam.es/redd2014.

4. REFERENCES

[1] Adamopoulos, P. and Tuzhilin, A. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM TIST*.

[2] Adomavicius, G. and Kwon, Y. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE TKDE*, 24(5):896 –911, 2012.

[3] Breese, J. S., Heckerman, D., and Kadie, C. Empirical analysis of predictive algorithms for collaborative filtering. *14th Conf. on Uncertainty in Artificial Intelligence (UAI 1998)*, 43-52, 1998.

[4] Celma, O. and Herrera, P. A New Approach to Evaluating Novel Recommendations. *2nd ACM International Conference on Recommender Systems (RecSys 2008)*, 179-186, 2008.

[5] Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A. V., and Turrin, R. Comparative evaluation of recommender system quality. *ACM Conference on Human Factors in Computing Systems (CHI 2011)*, 1927-1932, 2011.

[6] Cremonesi, P., Koren, Y., and Turrin, R. Performance of Recommender Algorithms on Top-N Recommendation Tasks. *4th ACM International Conference on Recommender Systems (RecSys 2010)*, 39-46, 2010.

[7] Fleder, D. M. and Hosanagar, K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science,* 55(5):697-712, 2009.

[8] Ge, M., Delgado-Battenfeld, C. and Jannach, D. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. *4th ACM International Conference on Recommender Systems (RecSys 2010)*, 257-260, 2010.

[9] Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. Evaluating collaborative filtering recommender systems. *ACM Trans. on Information Systems,* 22(1):5-53, 2004.

[10] Lathia, N., Hailes, S., Capra, L., and Amatriain, X. Temporal Diversity in Recommender Systems. *33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2010)*, 210-217, 2010.

[11] McNee, S. M., Riedl, J., and Konstan, J. A. Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems. *ACM Conf. on Human Factors in Computing Systems (CHI 2006)*, 1097-1101, 2006.

[12] Shani, G. and Gunawardana, A. Evaluating Recommendation Systems. In Ricci, F. et al (Eds.), *Recommender Systems Handbook*, pages 257-297. Springer, 2011.

[13] Steck, H. Training and Testing of Recommender Systems on Data Missing Not At Random. *16th ACM International Conference on Knowledge Discovery and Data Mining (KDD 2010)*, 713-722, 2010.

[14] Vargas, S. and Castells, P. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. *5th ACM International Conference on Recommender Systems (RecSys 2011)*, 109-116, 2011.

[15] Voorhees, E. M. and Harman, D. K. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.