Introduction to the Special Issue on Recommender System Benchmarking

PAOLO CREMONESI, Politecnico di Milano
ALAN SAID, Recorded Future
DOMONKOS TIKK, Gravity R&D
MICHELLE X. ZHOU, Juji

CCS Concepts: • **Information systems → Personalization; Retrieval models and ranking; Relevance assessment; Presentation of retrieval results;**

Additional Key Words and Phrases: Recommender systems, benchmarking, evaluation

ACM Reference Format:
Paolo Cremonesi, Alan Said, Domonkos Tikk, and Michelle X. Zhou. 2016. Introduction to the special issue on recommender system benchmarking. ACM Trans. Intell. Syst. Technol. 7, 3, Article 38 (March 2016), 4 pages.
DOI: http://dx.doi.org/10.1145/2870627

1. INTRODUCTION

Recommender systems addvalue to vast content resources by matching users with items of interest. In recent years, immense progress has been made in recommendation techniques. The evaluation of these systems is still based on traditional information retrieval and statistics metrics (e.g., precision, recall, RMSE), often not taking the use case and situation of the system into consideration. However, the rapid evolution of recommender systems in both their goals and their application domains fosters the need for new evaluation methodologies and environments. This special issue serves as a venue for work on novel, recommendation-centric benchmarking approaches taking the users' utility, the business values, and the technical constraints into consideration.

Building on the success of the Recommendation Utility Evaluation Workshop [Amatriain et al. 2012] held at Recsys 2012, the Workshop on Benchmarking Adaptive Retrieval and Recommender Systems [Castells et al. 2013a, 2013b] held at SIGIR 2013, the Workshop on Reproducibility and Replication in Recommender System Evaluation [Bellogín et al. 2013, 2014], the various Recommender System Challenges [Adomavicius et al. 2010; Said et al. 2011; Manouselis et al. 2012; Blomo et al. 2013; Said et al. 2014; Ben-Shimon et al. 2015], and other similar events, this special issue collects articles focusing on various aspects of challenges in benchmarking and evaluation or recommender systems.

In the decade since the inception of the ACM RecSys conference and the decades since the first papers on this topic started to appear [Resnick et al. 1994], the field has

Authors' addresses: P. Cremonesi, Departimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci ew, 20133, Milan, Italy; A. Said, Recorded Future, Västra Hamngatan 24, 411 17, Göteborg, Sweden; email: alansaid@acm.org; D. Tikk, Gravity R&D, 1016 Budapest, Meszaros u. 58/b, Hungary; email: domonkos.tikk@gravityrd.com; M. Zhou, Juji, P.O. Box 2907, Saratoga, CA, USA, 95070.

2157-6904/2016/03-ART38
DOI: http://dx.doi.org/10.1145/2870627

grown increasingly mature, and with this the focus on evaluation and benchmarking has increased.

2. ARTICLES

The special issue contains five articles on various aspects of recommender system benchmarking. An overview of each article is presented next.

Relevance Meets Coverage: A Unified Framework to Generate Diversified Recommendations. Wu et al. [2016] bring to light the competing aspects of recommending relevant items to the user versus the catalog coverage of the vendor. The authors show that traditional user-based collaborative filtering magnifies the popularity bias, and recommendations lack diversity. The overcome this, the article first introduces a simple measure of coverage that quantifies the usefulness of the neighbor userset and the recommended itemset as a complete entity, and then a recommendation framework REC (RElevance + Coverage) is proposed to generate diversified top-N recommendations, maximizing relevance and coverage measures simultaneously. Though both the neighbor set selection step and the recommendation set generation step are found to be NP-hard, they can be solved effectively and efficiently by exploiting the inherent submodular property. Their experiments performed on three real-world datasets show that the REC-based recommendation models can naturally generate more diversified recommendations without sacrificing the accuracy.

The Role of Cores in Recommender Benchmarking for Social Bookmarking Systems. Doerfel et al. [2016] focus on the benchmarking of tag recommender systems. They first show that the typical offline evaluation of tag recommender systems uses datasets that are pruned to the core datasets, and the influence of this restriction is not taken into account at benchmarking. The authors introduce the concepts of set-core to overcome certain structural drawbacks on tagging datasets and show that by using set-cores, some problems relevant to cores can be eliminated. The article provides a comprehensive benchmark of tag recommenders and core setups, and in a large-scale experiment they investigate, using four real-world datasets, the impact of different cores on the evaluation of the recommender algorithms. Their results suggest that comparison of different recommendation approaches depends on the selection of core type and level, and thus the evaluation setup must be prepared carefully to avoid biases.

A Framework for Dataset Benchmarking and Its Application to a New Movie Rating Dataset. Dooms et al. [2016] present an extensive analysis of a new, continuously updated, movie rating dataset. Movie rating datasets, like MovieLens and Netflix, were always of paramount importance in recommender systems research, but they are static and became old. The article proposes to combine new data sources, using the emerging trends of social media and smartphones, to create valuable research datasets. This work proposes a five-step framework to introduce and benchmark new datasets in the recommender systems domain. The framework is illustrated on a new movie rating dataset called MovieTweetings collected from Twitter. The authors analyze the new dataset from different aspects following their framework, from descriptive statistics, to investigate external validity, and they report on a number of reproducible benchmarks.

A Novel Classification Framework for Evaluating Individual and Aggregate Diversity in Top-N Recommendations. Moody and Glass [2016] propose a framework for user-centered evaluation of recommender systems. In this classification framework, user profiles are constructed and matched against other users' profiles to formulate neighborhoods and generate top-N recommendations. These recommendations are then evaluated for accuracy, coverage, and diversity for groups of users in the framework, using also some new diversity metrics introduced in the article. The article also suggests a

novel definition of sparse users in terms of low average similarity with nearest neighbors and low number of ratings to uncover the measured differences of recommendation evaluation on user groups. As opposed to common assumptions, the authors show that not all users suffer equally from the data sparsity problem, and interestingly, the group of users that receive the most accurate recommendations do not belong to the least sparse area of the dataset.

Anytime Algorithms for Recommendation Service Providers. Ben-Shimon et al. [2016] discuss the pros and cons of fast recommendations versus accurate recommendations from the perspective of a recommender system (RS) vendor. These companies must carefully balance the cost of building recommendation models and their revenues from clients, due to the competitive market. One factor that may influence the profitability of RS vendors is the computational time required for model building. In this commercially motivated setting, the authors propose anytime algorithms as an attractive tradeoff solution for balancing computational time and the recommendation model performance. In an anytime setting, an algorithm can be stopped after any amount of computational time, always ensuring that a valid, although suboptimal, solution will be returned. The recommendation quality of the returned solution should be monitored over time, which is a novel evaluation strategy. It is shown that the popular item-item top-N recommendation approach can be brought into the anytime framework, by smartly considering the order by which item pairs are being evaluated.

REFERENCES

Gediminas Adomavicius, Alexander Tuzhilin, Shlomo Berkovsky, Ernesto W. De Luca, and Alan Said. 2010. Context-awareness in recommender systems: Research workshop and movie recommendation challenge. In *Proceedings of the 4th ACM Conference on Recommender Systems (RecSys'10)*. ACM, New York, NY, 385–386. DOI:http://dx.doi.org/10.1145/1864708.1864801

Xavier Amatriain, Pablo Castells, Arjen de Vries, Christian Posse, and Harald Steck (Eds.). 2012. *Proc. of the Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE'12)*.

Alejandro Bellogín, Pablo Castells, Alan Said, and Domonkos Tikk. 2013. Workshop on reproducibility and replication in recommender systems evaluation: RepSys. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys'13)*. ACM, New York, NY, 485–486. DOI:http://dx.doi.org/10.1145/2507157.2508006

Alejandro Bellogín, Pablo Castells, Alan Said, and Domonkos Tikk. 2014. Report on the workshop on reproducibility and replication in recommender systems evaluation (RepSys). *SIGIR Forum* 48, 1 (June 2014), 29–35. DOI:http://dx.doi.org/10.1145/2641383.2641389

David Ben-Shimon, Lior Rokach, Guy Shani, and Bracha Shapira. 2016. Anytime algorithms for recommendation service providers. *ACM Trans. Intell. Syst. Technol.* (2016).

David Ben-Shimon, Alexander Tsikinovsky, Michael Friedmann, Bracha Shapira, Lior Rokach, and Johannes Hoerle. 2015. RecSys challenge 2015 and the YOOCHOOSE dataset. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys'15)*. ACM, New York, NY, 357–358. DOI:http://dx.doi.org/10.1145/2792838.2798723

Jim Blomo, Martin Ester, and Marty Field. 2013. RecSys challenge 2013. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys'13)*. ACM, New York, NY, 489–490. DOI:http://dx.doi.org/10.1145/2507157.2508008

Pablo Castells, Frank Hopfgartner, Alan Said, and Mounia Lalmas. 2013a. Report on the SIGIR 2013 workshop on benchmarking adaptive retrieval and recommender systems. *SIGIR Forum* 47, 2 (Jan. 2013), 64–67. DOI:http://dx.doi.org/10.1145/2568388.2568398

Pablo Castells, Frank Hopfgartner, Alan Said, and Mounia Lalmas. 2013b. Workshop on benchmarking adaptive retrieval and recommender systems: BARS 2013. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'13)*. ACM, New York, NY, 1133–1133. DOI:http://dx.doi.org/10.1145/2484028.2484224

Stephan Doerfel, Robert Jaeschke, and Gerd Stumme. 2016. The role of cores in recommender benchmarking for social bookmarking systems. *ACM Trans. Intell. Syst. Technol.* (2016).

Simon Dooms, Alejandro Bellogín, Toon De Pessemier, and Luc Martens. 2016. A framework for dataset benchmarking and its application to a new movie rating dataset. *ACM Trans. Intell. Syst. Technol.* (2016).

Nikos Manouselis, Alan Said, Domonkos Tikk, Jannis Hermanns, Benjamin Kille, Hendrik Drachsler, Katrien Verbert, and Kris Jack. 2012. Recommender systems challenge 2012. In *Proceedings of the 6th ACM Conference on Recommender Systems (RecSys'12)*. ACM, New York, NY, 353–354. DOI:http://dx.doi.org/10.1145/2365952.2366043

Jennifer Moody and David H. Glass. 2016. A novel classification framework for evaluating individual and aggregate diversity in top-N Recommendations. *ACM Trans. Intell. Syst. Technol.* (2016).

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW'94)*. ACM, New York, NY, 175–186. DOI:http://dx.doi.org/10.1145/192844.192905

Alan Said, Shlomo Berkovsky, Ernesto William De Luca, and Jannis Hermanns. 2011. Challenge on context-aware movie recommendation: CAMRa2011. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys'11)*. ACM, New York, NY, 385–386. DOI:http://dx.doi.org/10.1145/2043932.2044015

Alan Said, Simon Dooms, Babak Loni, and Domonkos Tikk. 2014. Recommender systems challenge 2014. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys'14)*. ACM, New York, NY, 387–388. DOI:http://dx.doi.org/10.1145/2645710.2645779

Le Wu, Qi Liu, Enhong Chen, Nicholas Jing Yuan, Guangming Guo, and Xing Xie. 2016. Relevance meets coverage: A unified framework to generate diversified recommendations. *ACM Trans. Intell. Syst. Technol.* (2016).

Received November 2015; revised December 2015; accepted December 2015