Proceedings of the

Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2011)

held at the

6th ACM International Conference on Recommender Systems (RecSys 2012)

9 September 2012

Dublin, Ireland

Edited by

Xavier Amatriain[1], Pablo Castells[2], Arjen de Vries[3],
Christian Posse[4], Harald Steck[1]

[1] Netflix, USA
[2] Universidad Autónoma de Madrid, Spain
[3] Centrum Wiskunde & Informatica, Netherlands
[4] Linkedin, USA

Preface

Introduction

Measuring the error in predicting held-out user rating values has been by far the dominant offline evaluation methodology in the Recommender Systems (RS) literature. Yet there seems to be a general consensus in the community that this criterion alone is far from being enough or even adequate to assess the practical effectiveness of a recommender system in matching user needs. The end users of recommendations receive lists of items rather than rating values, whereby recommendation accuracy metrics –as surrogates of the evaluated task– should target the quality of the item selection, rather than the numeric system scores that determine this selection. Furthermore, as far as the order of recommended items determines the set of elements that the user will actually consider for consumption, effectiveness assessment methodologies should target item rankings. For this reason, metrics and methodologies from the Information Retrieval (IR) field –where ranking evaluation has been studied and standardized for decades– have started to be adopted by the RS community. Gaps remain between the methodological formalization of tasks in both fields though, which result in divergences in the adoption of IR methodologies for RS, hindering the interpretation and comparability of empirical observations by different authors.

On the other hand, there is a growing realization that accuracy is only one among several relevant dimensions of recommendation effectiveness. The value of novelty, for instance, has been recognized as a key dimension of recommendation utility for users in real scenarios, in-as-much as the purpose of recommendation is inherently linked to discovery in many application domains. Closely related to novelty, diversity is also a desirable quality to enrich the user's experience and enhance his array of relevant choices. Novelty and diversity are generally positive for businesses as well, by favoring the diversity of sales and helping leverage revenues from market niches. As a matter of business performance enhancement, the value added by recommendation can be measured more directly in terms of on-line click-through rate, conversion rate, sales order size increase, returning customers, increased revenue, etc. On the other hand, web portals and social networks commonly face multiple objective optimization problems related to user engagement, requiring appropriate evaluation methodologies for optimizing along the entire recommendation funnel, from the initial click to the real user engagement in subsequent downstream utilities. Other potentially relevant dimensions of effective recommendations for consumers and providers may include confidence, coverage, risk, cost, robustness, etc.

While the need for further extension, formalization, clarification and standardization of evaluation methodologies is recognized in the community, this need is still unmet to a large extent. When engaging in evaluation work, researchers and practitioners are still often faced with experimental design questions for which there are currently not always precise and consensual answers. Room re-mains for further methodological development and convergence, which motivated the RUE 2012 workshop

The ACM RecSys 2012 International Workshop on "Recommendation Utility Evaluation: Beyond RMSE" (RUE 2012) gathered researchers and practitioners interested in developing better, clearer, and/or more complete evaluation methodologies for recommender systems –or just seeking clear guidelines for their experimental needs. The workshop provided an informal setting for exchanging and discussing ideas, sharing experiences and viewpoints. RUE sought to identify and better understand the current gaps in recommender system evaluation methodologies, help lay directions for progress in addressing them, and contribute to the consolidation and convergence of experimental methods and practice.

Scope and topics

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

- Recommendation quality dimensions.
  - Effective accuracy, ranking quality.
  - Novelty, diversity, unexpectedness, serendipity.
  - Utility, gain, cost, risk, benefit.
  - Robustness, confidence, coverage, usability, etc.
- Matching metrics to tasks, needs, and goals.
  - User satisfaction, user perception, human factors.
  - Business-oriented evaluation.
  - Multiple objective optimization, user engagement.
  - Quality of service, quality of experience.
- Evaluation methodology and experimental design.
  - Definition and evaluation of new metrics, studies of existing ones.
  - Adaptation of methodologies from related fields: IR, Machine Learning, HCI, etc.
  - Evaluation theory.
- Practical aspects of evaluation.
  - Offline and online experimental approaches.
  - Simulation-based evaluation.
  - Datasets and benchmarks.
  - Validation of metrics.

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

- What are the unmet needs and challenges for evaluation in the RS field? What changes would we like to see? How could we speed up progress?
- What relevant recommendation utility and quality dimensions should be cared for? How can they be captured and measured?
- How can metrics be 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 disciplines should we draw from (HCI, Machine Learning, etc.)?
- What biases and noise should experimental design typically watch for?
- Can we predict the success of a recommendation algorithm with our 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 do we envision the evaluation of recommender systems in the future?

Submissions and Programme

The workshop received 18 submissions, of which 11 were accepted (61%), including 3 full technical papers, 4 position papers, and 4 technical papers presented as posters. The workshop opened with a keynote talk by Carlos Gómez-Uribe, from Netflix, and included several open discussion sessions. We briefly summarize here the presented works and held discussions.

The keynote talk, entitled "Challenges and Limitations in the Offline and Online Evaluation of Recommender Systems: A Netflix Case Study", provided a comprehensive, inside view of the evaluation of recommendation technologies in one of the major players in the recommender system industry. Gómez-Uribe explained and discussed how offline and online (A/B testing) phases, business metrics (cancellation rate, subscriber streaming), long-term vs. short-term performance measures are handled in an online businesses heavily relying on recommendation technologies.

The papers presented after this cover a wide spectrum of topics, encompassing most of the aspects put forward in the intended workshop scope. In the full technical papers section, G. Adomavicius and J. Zhang address a new quality dimension, namely recommendation stability, defined as the consistency of recommendations over small incremental changes in the input data. A method is proposed to enhance the stability (at the same time as the accuracy) of an arbitrary recommender by means of an iterative approach where the system is fed back samples of its output. F. Meyer et al. propose the distinction of four functions in user activity where a recommender system may assist: decision, comparison, discovery, and exploration. The authors suggest associating a specific evaluation metric to each of these dimensions, and a structured evaluation procedure (including the metrics computation) in offline experiments. M. Habibi and A. Popescu-Belis present a crowdsourcing approach to evaluate the accuracy of a filtering system which automatically links documents to human speech. The study addresses such issues as worker's reliability assessment, inter-worker agreement, and evaluation stability.

In the position papers section, A. Said, D. Tikk et al. present a conceptual framework where evaluation considers three dimensions: the business model, the user requirements, and technical constraints, corresponding to the view of the three broad types of stakeholders involved in a recommender application, respectively: vendors, consumers, and service providers. S. Clerger-Tamayo, J. M. Fernández-Luna and J. F. Huete propose a generalization of MAE where the error in rating predictions can be weighted in order to focus the evaluation on specific cases of the user-item space, and identify conditions where recommendation is suboptimal. O. Başkaya and T. Aytekin study the correspondence between rating-based and content-based inter-item similarity, which is a relevant issue for metrics that are based on a generic item similarity function (such as the average intra-list dissimilarity for diversity evaluation). B. Kille considers the fact that different users may not be equally easy to provide accurate recommendations for, and proposes measures to assess user difficulty in this perspective.

In the posters section, W. L. De Mello Neto and A. Nowé contrast offline and online evaluation in the context of recommendation approaches leveraging social network information, considering such issues as transparency and computational complexity, besides recommendation accuracy. K. Oku and F. Hattori present an approach to enhance recommendation serendipity by mixing features from different items as a seed to produce new recommendations. C. E. Seminario and D. C. Wilson report a wide empirical study with the Mahout open source library, focusing on accuracy and coverage, the tradeoffs between such dimensions, and the variations resulting from functional enhancements introduced by the authors. L. Peska and P. Vojtas present an experimental study on a travel agency website, where the extended use of implicit evidence from user interaction as input for recommendation is tested, using clickthrough rate and conversion rate as the primary recommendation performance metrics.

Different issues were addressed in the open discussion sessions during the workshop. A prominently recurrent one was the gap between offline and online evaluation. Academic research is strongly focused on offline experiments using rating prediction error or IR metrics, whereas businesses rely on live A/B testing with real customers using business metrics such as CTR, conversion rate, cancellation rate (equivalently, returning customers), or revenue increase (in its different measurable forms). With sales and profitability as the obvious common baseline denominator, the designation of specific core metrics among these seems use-case dependent. Some businesses, such as Netflix, nonetheless report using offline testing as well, as a preliminary phase prior to online experimentation. Business-oriented evaluation goals typically require longer-term evaluation cycles, where the effects of a feature (e.g. on customer loyalty) can only be measured over an extended period of time (several months). Short-term indicators (such as CTR) are commonly used nonetheless to complement these to some extent.

Some of the pointed out hurdles hindering the connection between academia and industry in this area include the often discussed difficulty for academic researchers to access real-world large-scale datasets, or the availability of a feasible procedure where algorithms from academia and data from vendors might get in contact, while meeting the requirements and constraints of the involved stakeholders (data and algorithm ownership, end-user's privacy, etc.). Public evaluation campaigns such as the Netflix Prize, the CAMRa challenges, the plista Contest, were discussed as very positive moves in this direction, each with their own limitations. Further alternatives were discussed for setting up some form of evaluation platform where systems could be compared not just for accuracy, but for scalability (response time). Paolo Cremonesi discussed also an initiative, currently in perspective, which is aiming in this direction. The open API provided by Mendeley to its data was also described as an available opportunity for researchers to test their algorithms on massive data. Crowdsourcing approaches were furthermore mentioned as an intermediate option, available to researchers, between offline evaluation and full-scale experiments on real application data.

There seemed to be a general consensus on the inadequacy of RMSE as a proxy for user satisfaction, or any proper view on recommendation utility in general. This was expressed from many perspectives: from conceptual rationales (the recommender's task, the user's goals and interaction paradigm with recommendations in real applications) to experiences –offline and online evidence in formal or informal case studies– shared by many participants in the workshop. Several other general concerns were identified regarding the adequacy of different metrics, such as the fact that different contexts may require different metrics, e.g. navigational recommendation may focus on accuracy and diversity, whereas discovery-oriented recommendation may emphasize novelty and serendipity. It is also a common case in industrial contexts that technical requirements and business constraints may have to be traded off with evaluation needs and may override other concerns and observations in A/B testing. Beyond RMSE, the general opinion seemed yet to be that researchers in academia should still focus on generic metrics rather than too specific business-oriented metrics and constraints.

The interest of the workshop theme was underlined, beyond the RUE workshop itself, by the pervading presence of evaluation as an explicit object of research and discussion in the RecSys conference programme, clearly identified as an open area where further work is needed.

Acknowledgments

The organizers would like to thank the Program Committee members for their high-quality and timely evaluation of the submissions; the RecSys 2012 organizers (Neil J. Hurley and his team) and workshop chairs (Jill Freyne and Pearl Pu) for their support in the organization of this workshop; the keynote speaker (Carlos Gómez-Uribe), all the authors and presenters, for their contribution to a high-quality workshop program; and all participants for such fruitful discussions and valuable ideas as were exchanged during the workshop. Thanks are due to all such contributions which made RUE 2012 a successful venue.

Xavier Amatriain
Pablo Castells
Arjen de Vries
Christian Posse
Harald Steck

Organizing Committee

Xavier Amatriain	Netflix, USA
Pablo Castells	Universidad Autónoma de Madrid, Spain
Arjen de Vries	Centrum Wiskunde & Informatica, Netherlands
Christian Posse	Linkedin, USA
Harald Steck	Netflix, USA

Program Committee

Gediminas Adomavicius	University of Minnesota, USA
Alejandro Bellogín	Universidad Autónoma de Madrid, Spain
Iván Cantador	Universidad Autónoma de Madrid, Spain
Licia Capra	University College London, UK
Òscar Celma	Gracenote, USA
Charles Clarke	University of Waterloo, Canada
Paolo Cremonesi	Politecnico di Milano, Italy
Juan Manuel Fernández-Luna	Universidad de Granada, Spain
Pankaj Gupta	Twitter, USA
Juan F. Huete	Universidad de Granada, Spain
Dietmar Jannach	University of Dortmund, Germany
Jaap Kamps	University of Amsterdam, Netherlands
Neal Lathia	University College London, UK
Jérôme Picault	Bell Labs, Alcatel-Lucent, France
Filip Radlinski	Microsoft, Canada
Francesco Ricci	Free University of Bozen-Bolzano, Italy
Fabrizio Silvestri	Consiglio Nazionale delle Ricerche, Italy
David Vallet	Universidad Autónoma de Madrid, Spain
Paulo Villegas	Telefónica R&D, Spain
Jun Wang	University College London, UK
Yi Zhang	University of California, Santa Cruz, USA

Table of Contents