# Information Retrieval and User-Centric Recommender System Evaluation

Alan Said[†], Alejandro Bellogín[†], Arjen De Vries[†], Benjamin Kille[⋆]

CWI, The Netherlands[†], TU Berlin, Germany[⋆]
{alan, alejandro.bellogin, arjen.de.vries}@cwi.nl[†], kille@dai-lab.de[⋆]

**Abstract.** Traditional recommender system evaluation focuses on raising the accuracy, or lowering the rating prediction error of the recommendation algorithm. Recently, however, discrepancies between commonly used metrics (e.g. precision, recall, root-mean-square error) and the experienced quality from the users' have been brought to light. This project aims to address these discrepancies by attempting to develop novel means of recommender systems evaluation which encompasses qualities identified through traditional evaluation metrics and user-centric factors, e.g. diversity, serendipity, novelty, etc., as well as bringing further insights in the topic by analyzing and translating the problem of evaluation from an Information Retrieval perspective.

**Keywords:** Recommender Systems, Evaluation, Information Retrieval

## 1 Introduction

The project is framed in the Recommender Systems (RS) field. The aim of RSs is to assist users in finding their way through huge databases and catalogues, by filtering and suggesting relevant items, taking into account the users' preferences (i.e., tastes, priorities, etc.).

## 2 Novel Methods for Recommender System Evaluation

Over the last two decades, a vast amount of research in RS has lead to great progress in terms of prediction accuracy [8]. Today, the majority of the work on RS is based on top-n recommendation or rating prediction; the former requires bi/unary interaction data between users and items, whereas the latter requires a dataset with ratings. This type of evaluation is also common in information retrieval (IR) systems [2].

### 2.1 User-centric Evaluation

Both top-n and rating prediction-based evaluation build on several assumptions which could potentially have negative effects on recommendation algorithms tuned solely on these evaluation metrics [1, 7–9]. These are:

– there is an absolute ground truth which the RS should attempt to identify,
– users are primarily interested in the items which have received the highest ratings,
– higher top-n accuracy or lower rating error levels translate to a higher perceived usefulness from the users.

Recent work has, however, shown that these assumptions are not always true in the RS context, e.g. [10–12]. In specific cases, the assumptions are detrimental to the users' perceived quality.

The main focus of this sub-project is to analyze the discrepancies between *offline* and *online* evaluation, and to gain insights into the subjective qualities of various recommendation algorithms and their specific qualities. With this in mind, there are several goals we will strive to achieve:

– Analyze the correlation of IR-related evaluation metrics and user-centric concepts such as diversity, novelty and other non-quantifiable RS aspects.
– Identify whether there exists a correlation between the properties of items that are regarded as false recommendations in traditional (offline) IR and RS evaluation settings and true (high quality) recommendations in user-centric (online) evaluation, taking factors such as serendipity and usefulness into consideration.
– Evaluate whether metrics from research areas outside of IR and RS (e.g. signal processing, economics) can estimate sought for qualities better than the currently used RS and IR metrics.
– Improve the understanding of which evaluation metrics should be applied to RSs in different contexts, using different algorithms, data sets, and other system-specific features.

For these purposes, we will use a variety of data sets from the multimedia domain, ranging from publicly available, e.g. Movielens,Last.fm, as well as proprietary from other related services, e.g. Filmtipset, Moviepilot, etc. The variety of data will ensure the general applicability of the research results. Additional data will be collected through user studies and surveys.

### 2.2 Information Retrieval-based Evaluation

RS are usually considered as a special case of IR systems, specifically, one where no query is given and the information to be retrieved has to be inferred from previous user experiences. For this reason, some of the models and theories developed in IR have already been translated to RS, such as the Vector Space Model and the Probability Ranking Principle [13].

There are, however, several gaps in the understanding of RS as personalized IR systems, such as the need of formal methods to introduce implicit and contextual feedback in the recommendations and the lack of a proper evaluation framework.

A strong link between IR and RS has already been shown in our previous research (see [3, 6]), where we adapted to RSs different techniques proposed in

IR to predict the performance of a system. A natural next step is to explore how the evaluation in RS may benefit from extending this analogy between IR and recommendation, and applying more retrieval methodologies to recommendation.

More specifically, in this sub-project, we aim to exploit IR concepts, algorithms, and methodologies for recommendation .RS has a well known tradition in integrating contextual information which could be, in turn, transferred from RS to IR and investigate how such methods could be integrated in contextual IR and whether some benefits could be found by creating such links.

With this goal in mind, there are several objectives we aim to achieve:

- Analyze how evaluation in recommendation should be performed to obtain a general methodology that would result in interpretable and comparable results for the community, mainly by adapting and integrating models and metrics from IR.
- Identify if there is any correlation between the metrics typically used in offline experiments (e.g., precision) and those pervasive in real applications, more useful from a business point of view (such as the click through and conversion rates).
- Bring the models and theories used in context-aware recommendation to contextual IR and vice versa, in such a way that further interactions between these two areas could be found.
- Exploit implicit information for RS in novel ways based on current research from the use of search logs and other implicit sources of information in IR.

## 3 Current Results

At the moment of writing, we have obtained positive results regarding some of the aspects of this research project. Specifically, a workshop on reproducibility and replication in recommender evaluation has been accepted at the 2013 ACM RecSys conference. We are also working on revisiting several classic recommendation techniques in order to evaluate them under a common framework based on an IR-inspired recommendation method previously proposed in [5]. A novel evaluation protocol has also been researched specifically for RS. Additionally, we are researching different features that could help in the understanding of why some similarity functions perform better when applied within a recommendation strategy. Finally, we have analyzed different sources of implicit and explicit popularity scores in order to exploit them in a recommendation context; this study was recently accepted for publication as [4].

## 4 Acknowledgements

This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Programme. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no.246016.

## References


1. Amatriain, X., Pujol, J.M., Oliver, N.: I like it... i like it not: Evaluating user ratings noise in recommender systems. In: Proceedings of the 17th International Conference on User Modeling, Adaptation, and Personalization: formerly UM and AH. pp. 247–258. UMAP '09, Springer-Verlag, Berlin, Heidelberg (2009), http://dx.doi.org/10.1007/978-3-642-02247-0_24
2. Baeza-Yates, R.A., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1999)
3. Bellogín, A.: Recommender System Performance Evaluation and Prediction: an Information Retrieval Perspective. Ph.D. thesis, Universidad Autonóma de Madrid, Spain (Nov 2012)
4. Bellogín, A., de Vries, A., He, J.: Artist popularity: do web and social music services agree? In: Int. Conf. on Weblogs and Social Media (ICWSM). Boston (2013)
5. Bellogín, A., Wang, J., Castells, P.: Bridging memory-based collaborative filtering and text retrieval. Information Retrieval pp. 1–28 (2012), http://dx.doi.org/10.1007/s10791-012-9214-z
6. Bellogn, A.: Performance prediction in recommender systems. In: Konstan, J., Conejo, R., Marzo, J., Oliver, N. (eds.) User Modeling, Adaption and Personalization, Lecture Notes in Computer Science, vol. 6787, pp. 401–404. Springer Berlin Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-22362-4_37
7. Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A., Turrin, R.: Comparative evaluation of recommender system quality. In: Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems. pp. 1927–1932. CHI EA '11, ACM, New York, NY, USA (2011), http://doi.acm.org/10.1145/1979742.1979896
8. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. 22(1), 5–53 (Jan 2004), http://doi.acm.org/10.1145/963770.963772
9. Hill, W., Stead, L., Rosenstein, M., Furnas, G.: Recommending and evaluating choices in a virtual community of use. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 194–201. CHI '95, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (1995), http://dx.doi.org/10.1145/223904.223929
10. Said, A., Fields, B., Jain, B.J.: User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In: Proceedings of the ACM 2013 conference on Computer Supported Cooperative Work. ACM, New York, NY, USA (2013)
11. Said, A., Jain, B., Narr, S., Plumbaum, T.: Users and noise: The magic barrier of recommender systems. In: Masthoff, J., Mobasher, B., Desmarais, M., Nkambou, R. (eds.) User Modeling, Adaptation, and Personalization, Lecture Notes in Computer Science, vol. 7379, pp. 237–248. Springer Berlin / Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-31454-4_20, 10.1007/978-3-642-31454-4_20
12. Said, A., Jain, B.J., Narr, S., Plumbaum, T., Albayrak, S., Scheel, C.: Estimating the magic barrier of recommender systems: a user study. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp. 1061–1062. SIGIR '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2348283.2348469
13. Wang, J., Robertson, S., Vries, A.P., Reinders, M.J.: Probabilistic relevance ranking for collaborative filtering. Inf. Retr. 11(6), 477–497 (Dec 2008), http://dx.doi.org/10.1007/s10791-008-9060-1