New approaches for evaluation: correctness and freshness

Extended Abstract

Pablo Sánchez
Universidad Autónoma de Madrid
Madrid, Spain
pablo.sanchezp@predoc.uam.es

Rus M. Mesas
Telefónica
Madrid, Spain
rusmaria.mesasjavega@telefonica.com

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

ABSTRACT

The main goal of a Recommender System is to suggest relevant items to users, although other utility dimensions – such as diversity, novelty, confidence, possibility of providing explanations – are often considered. In this work, we study two dimensions that have been neglected so far in the literature: coverage and temporal novelty. On the one hand, we present a family of metrics that combine precision and coverage in a principled manner (correctness); on the other hand, we provide a measure to account for how much a system is promoting fresh items in its recommendations (freshness). Empirical results show the usefulness of these new metrics to capture more nuances of the recommendation quality.

ACM Reference Format:
Pablo Sánchez, Rus M. Mesas, and Alejandro Bellogín. 2018. New approaches for evaluation: correctness and freshness: Extended Abstract. In *CERI 18: 5th Spanish Conference in Information Retrieval, June 26–27, 2018, Zaragoza, Spain.* ACM, New York, NY, USA, Article 14, 2 pages. https://doi.org/10.1145/3230599.3230614

1 INTRODUCTION

Recommender Systems (RS) have become necessary applications in a large number of companies that offer personalized content to users. The ability to not only get useful and relevant recommendations, but also to offer novel and diverse items, can increase the number of users of a system and generate more benefits for companies. However, even though most recommenders are optimized to make accurate recommendations, nowadays it is recognized that other evaluation dimensions besides accuracy should be considered to properly model the user needs and understand why she wants a recommendation [5]; serendipity, novelty, and diversity, among others, are some of the criteria that are starting to get attention in the RS community beyond accuracy metrics [3].

In this extended abstract, we address two open issues present in current RS evaluation. First, we study how to balance recommendation coverage and typical accuracy metrics like precision (*correctness*, presented in [6]); such a tradeoff might be critical, since it is possible to achieve a very high precision recommending only one item to a unique user, which would result in a useless recommender. Second, we present a novelty metric for time-aware scenarios (*freshness*, from [8]), where most of the work so far has focused on how evaluation methodologies (how the data partitioning should be made and which items should be considered when generating the rankings) should be applied to the recommendation data [2]. We validate such measurements using well-known recommendation algorithms on different real-world datasets.

2 RS EVALUATION BASED ON CORRECTNESS

Typical ranking-based metrics – such as precision – assume that not returning an item which was previously asked to predict a rating for, is an advocate of that item being considered as not relevant by a specific recommendation method. However, this is in contrast with the (desired) situation that a recommender may not provide suggestions in some situations due to a low confidence in the accuracy of such predictions [3, 4]. We present an evaluation metric – defined in [6] – that is able to assess when a recommender decides not to recommend a specific item. To do this, we adapt an extension of accuracy proposed in the context of Question Answering by Peñas and Rodrigo in [7]. In that work, the authors assume that there are several questions to be answered by a system, each question has several options, but one (and only one) of those options is correct. If it is possible to give no response for a given question, this action should not be correct, but not incorrect either. Hence, the authors propose a general formulation giving a weight – proportional to the number of correctly answered questions – to the value of unanswered questions.

To apply this evaluation metric to recommendation, we assume that the set of recommenders we want to compare will receive the same list of items to be ranked, a standard situation shared by many evaluation methodologies [1]. Then, the equivalence between a Question Answering system and a recommender is made – in a user basis – by considering each recommendation algorithm as a different system that will answer (or not) to the questions available, represented as the candidate items to be ranked by a specific methodology. We instantiate four versions of this metric, two of them based on users (UserCorrectness and RecallUserCorrectness) and two for items (ItemCorrectness and RecallItemCorrectness), that measure the number of correct recommendations received in a user or item basis, but rewarding not recommending instead of recommending something incorrect; additionally, the recall versions of the metrics include the assumption that it is worse to not recommend items when there are still relevant items available.

3 RS EVALUATION BASED ON FRESHNESS

The evaluation of recommender systems has overlooked the temporal dimension, and most of the work has focused on how different

CERI 18, June 26–27, 2018, Zaragoza, Spain

ACM ISBN 978-1-4503-6543-7/18/06…$15.00
https://doi.org/10.1145/3230599.3230614

evaluation methodologies should be applied to the recommendation data [2], leaving aside the definition of evaluation metrics specifically tailored to the problem of time-aware recommendation. Our proposed time-aware novelty metrics extend the traditional novelty metrics for RS by integrating the time dimension of user-item interactions with the system. Based on the generic novelty and diversity framework presented in [9], we define four time-aware item novelty models $\text{nov}(i \mid \theta_t)$ that are integrated seamlessly in this framework. The core idea there is how to define the item novelty model $\text{nov}(i \mid \theta)$, where θ stands for a generic contextual variable. Thus, depending on how these item novelty models are defined, different novelty metrics can be formulated. For instance, by taking the complement of the probability that the item was seen, the Expected Popularity Complement (EPC) [9] novelty metric is obtained, or when a distance-based measure is used, we obtain the intra-list diversity metric (ILD) [10].

Therefore, in order to model a time-aware novelty metric, we propose to encode the time model of the items in the θ variable based on the timestamps of the system interactions with the items. Hence, these models measure the item novelty either based on the first appearance of that item in the system (**FIN**, from *First Item Novelty*) or depending on its closeness to the end of the training split, in other terms, near to the test split (**LIN**, *Last Item Novelty*). Additionally, we also define an item novelty by computing the average (**AIN**) or the median (**MIN**) of the item interactions. Finally, as a last step, we normalise these values with a min-max normalisation so that they could be used in the probabilistic framework presented in [9].

4 EXPERIMENTAL RESULTS

4.1 Correctness

We evaluated the four instantiations of the correctness metric in three real-world datasets (Jester and two versions of MovieLens), comparing a user-based nearest neighbour algorithm and a probabilistic matrix factorisation technique. We found that our metrics are less sensitive to the values being combined (precision, coverage, and recall) than other combination metrics such as the harmonic or the geometric mean, which also need to specify some parameters. More importantly, our metrics help to interpret the comparisons between the recommenders, since they reward unanswered recommendations above incorrect recommendations.

One important issue with some of these variations is that they tend to return very small values, which may make more difficult to interpret its values by a system designer. In particular, among the four variations of the correctness metric, ItemCorrectness returns especially very low values because it is normalised by the number of users in the system and, hence, it might be less useful from a practical point of view.

4.2 Freshness

The four temporal novelty models were evaluated on three real-world datasets (Epinions, MovieLens, MovieTweetings), using a temporal split and a pool of 12 recommendation algorithms, including near-optimal skylines and more simple methods. Our results indicate that some time-aware algorithms tend to return more novel items than other methods, however this is not always the case, since a recommender might depend on the temporal dimension but not be optimised to return more fresh items (for example, sequential recommenders focus on predicting the next item to be consumed but that may not be a recent one); it is not possible to measure this information by means of standard metrics.

Additionally, we observe that depending on the item novelty model used to compute the freshness of a recommendation, some differences arise. For example, LIN always produces very high values, which does not help to discriminate between the recommenders. Regarding the FIN model, we believe it might not be very useful in datasets where several items appear at the very beginning, since it would not discriminate those cases, just like the LIN model. Nevertheless even with these simple metrics we are able to explain the behaviour of some of the recommenders. On the other hand, the two models that aggregate all the interactions received by an item (AIN and MIN) are more robust to these situations, and, in particular, the one based on the median (MIN) is expected to be more robust to outliers by definition.

5 CONCLUSIONS

Recommender systems evaluation is an active field where several issues remain open. In this work we have summarised two papers where we deal with balancing precision and coverage [6] and where we considered the temporal dimension for novelty metrics [8]. These metrics have some advantages with respect to the classical metrics: correctness allows to account for the assumption that it is better to not provide a recommendation rather than recommending something not relevant, which might be very useful in some scenarios such as finances or insurance; on the other hand, freshness allows us to measure if a recommendation technique is prone to return fresh items or not. Our results show that, while the proposed metrics work as expected, they also open the possibility to be affected by some biases in the data that are not necessarily considered when measuring accuracy-based metrics.

ACKNOWLEDGMENTS

This work was supported by the research project TIN2016-80630-P (MINECO).

REFERENCES

[1] Alejandro Bellogín, Pablo Castells, and Iván Cantador. 2011. Precision-oriented evaluation of recommender systems: an algorithmic comparison. In *RecSys*. ACM, 333–336.

[2] Pedro G. Campos, Fernando Díez, and Iván Cantador. 2014. Time-aware recommender systems: a comprehensive survey and analysis of existing evaluation protocols. *User Model. User-Adapt. Interact.* 24, 1-2 (2014), 67–119.

[3] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook*. Springer, 265–308.

[4] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.* 22, 1 (2004), 5–53.

[5] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Making recommendations better: an analytic model for human-recommender interaction. In *CHI Extended Abstracts*. ACM, 1103–1108.

[6] Rus M. Mesas and Alejandro Bellogín. 2017. Evaluating Decision-Aware Recommender Systems. In *RecSys*. ACM, 74–78.

[7] Anselmo Peñas and Álvaro Rodrigo. 2011. A Simple Measure to Assess Non-response. In *ACL*. The Association for Computer Linguistics, 1415–1424.

[8] Pablo Sánchez and Alejandro Bellogín. 2018. Time-Aware Novelty Metrics for Recommender Systems. In *ECIR (Lecture Notes in Computer Science)*, Vol. 10772. Springer, 357–370.

[9] Saul Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *RecSys*. ACM, 109–116.

[10] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *WWW*. ACM, 22–32.