From the PRP to the Low Prior Discovery Recall Principle for Recommender Systems

Rocío Cañamares
Universidad Autónoma de Madrid
rocio.cannamares@uam.es

Pablo Castells
Universidad Autónoma de Madrid
pablo.castells@uam.es

ABSTRACT

We revisit the Probability Ranking Principle in the context of recommender systems. We find a key difference in the retrieval protocol with respect to query-based search, that leads to the identification of different optimal ranking principles for discovery-oriented recommendation. Based on this finding, we revise the effectiveness of common non-personalized ranking functions in respect to the new principles. We run an experiment confirming and illustrating our theoretical analysis, and providing further observations and hints for reflection and future research.

KEYWORDS

recommender systems, probability ranking principle, novelty, discovery, accuracy, evaluation, popularity

1 INTRODUCTION

Robertson [8] put forward and discussed the Probability Ranking Principle (PRP) stating that under certain assumptions, the optimal ranking for a given information need is by decreasing probability of relevance of the documents to the information need. Robertson described and analyzed cases where the PRP may fail, and potential restatements of the principle in view of such limitations. A profuse line of research followed up extending or reexamining the PRP, seeking better, more complete, or more generalized principles [11], or aiming to fit the particularities of specific IR scenarios (such as interactive retrieval [6] or multimedia retrieval [10], to name a few). The PRP remains nonetheless a prominent notion today at the foundation of IR theory.

In this paper we analyze the recommendation task [1,5] as a new use case in the spirit of this long strand of research, seeking and analyzing the definition of an optimal ranking, through a formal methodological approach. A particularity of recommendation compared to the search task is that item relevance is understood to be a fully personal and subjective matter, solely defined by each end-user's personal taste, whereas judging the relevance of a search result has a (non-null but) narrower scope for disagreement, limited by a specific information need and its explicit expression as a query. Yet the PRP analysis in the context of search [8] has similarly considered degrees of user-level subjectivity or disagreement (in particular, as a challenge to the PRP), whereby our present research can be connected to such prior work in more than one way.

An additional singularity in the recommendation task is that, in its most widespread statement, the system should avoid recommending items the target user has already been observed interacting with. This restriction applies in scenarios where the added-value of recommendation is tightly linked to a purpose of discovery, as a complement of what users can already have experienced by themselves, and the assistance that other information retrieval technologies such as search engines already provide. In terms of an evaluation experiment, the condition means that items with an observed interaction record for the target user should be excluded from the ranking delivered to this user. This restriction substantially changes the frame for the optimal ranking analysis, as we shall see.

SIGIR '18, July 2018, Ann Arbor, MI, USA

2 BASIC CONCEPTS AND NOTATION

The recommendation task considers a set of users $\mathcal{U}$, a set of items $\mathcal{I}$, and a set of observed interaction records between users and items that can be interpreted as evidence of the user liking or disliking the item (i.e. relevance or non-relevance). As a widespread simplification, we may assume interaction data consist of a binary value $r: \Omega \subset \mathcal{U} \times \mathcal{I} \to \{0,1\}$ so that $r(u,i) = 1$ if the user $u \in \mathcal{U}$ likes the item $i \in \mathcal{I}$, and $r(u,i) = 0$ otherwise. Following common terminology, we shall refer to $r(u,i)$ as a *rating*. Ratings are available only for a subset Ω (typically a tiny fraction) of all user-item pairs $\mathcal{U} \times \mathcal{I}$ –there would otherwise not be any recommendation task to solve.

Taking the available rating data as input, the task of a recommender system is to compute a score for all user-item pairs where a rating is missing, and thus generate a ranking of unrated items to be delivered to each user in the system. The system output is evaluated using further user ratings on the recommended items, to be collected somehow, taken as relevance judgments. Such judgments can be obtained in different ways, depending on the evaluation setting. For instance, in offline evaluation, judgments are sampled (as so-called test data) from the available rating dataset itself, hiding them from the recommender system to be evaluated, while the remaining ratings are supplied as input training data to the system. In our theoretical analysis we will to some extent abstract ourselves from the problem of obtaining judgments, and assume we will manage somehow to get the relevance information we need.

For the convenience of our formal analysis, we shall introduce two binary random variables $rated: \mathcal{U} \times \mathcal{I} \to \{0,1\}$ and $rel: \mathcal{U} \times \mathcal{I} \to \{0,1\}$, where $rated = 1$ iff a rating (be it positive or negative) by the user on the item is present in the input data, and $rel = 1$ iff the user likes the item, regardless of whether this is known to the system (by the presence of a rating) or not. With this notation we can express well-defined distributions, e.g. $p(rated|i)$ is the ratio of users in $\mathcal{U}$ who have rated item i, and $p(rel|i)$ is the fraction of users who like the item.

3 EXPECTED AND OPTIMAL PRECISION

Whereas Robertson [8] considered a variety of evaluation metrics and cutoffs in his analysis, we shall focus here on $P@1$ as a simplest metric to make our analysis more tractable. Given a recommendation for a user, $P@1$ is equal to 1 if the target user likes the top

ranked item, and 0 if she does not. The expectation of $P@1$ for a given recommendation R is hence $\mathbb{E}[P@1|R] = p(P@1 = 1|R)$.

Now we need to be more precise with the computation of the metric: $P@1 = 1$ if the first ranked *recommendable* item in R is relevant. Let this item be i_k, ranked in the k-th position of R. As stated in the introduction, recommendable means that i_k does not have a rating by the target user, and being the first means that all the items $i_1, i_2, \ldots, i_{k-1}$ above i_k in R are not recommendable because they do have a rating. If we marginalize $p(P@1 = 1|R)$ by the possibility that the k-th item is the first recommendable, and we make the mild assumption that whether two items are rated or not by some user are mutually independent events, we have:

$$\mathbb{E}[P@1|R] \sim \sum_{k=1}^{|\mathcal{I}|} p(rel, \neg rated|i_k) \prod_{j=1}^{k-1} p(rated|i_j) \quad (1)$$

We should note how this equation contrasts with not considering item exclusion, in which case we would simply have $\mathbb{E}[P@1|R] \sim p(rel|i_1)$ and $\mathbb{E}[P@n|R] \sim \sum_{k=1}^{n} p(rel|i_k)$ as in [7], and the PRP analysis would be similarly applicable here. The exclusion of rated items can thus make a difference in the metric and, potentially, in the outcome of a comparative evaluation of algorithms.

3.1 Discovery False Negative Principle

We can now set forth the following result on the optimal non-personalized ranking for expected precision.

Lemma. Assuming pairwise item rating independence, the optimal recommendation that maximizes the expected $P@1$ ranks items $i \in \mathcal{I}$ by non-increasing value of $p(rel|\neg rated, i)$.

Proof. It suffices to show that a swap against $p(rel|\neg rated, i)$ in a ranking produces a smaller value for the expected $P@1$. Given that any ranking can be generated by a sequence of pairwise swaps on any other ranking (as per e.g. the proof of correction of bubble sort), we would have proven our point.

Let $R = \langle i_1, \ldots, i_n \rangle$ be some ranking so that $p(rel|\neg rated, i_k) \geq p(rel|\neg rated, i_{k+1})$ for some k, and let us consider a ranking R' consisting of swapping i_k and i_{k+1} in R. Using equation 1 it is easy to see that, by trivial algebraic cancellation and rearrangement of terms, we have:

$$\mathbb{E}[P@1|R] \geq \mathbb{E}[P@1|R'] \Leftrightarrow \frac{p(rel, \neg rated|i_k)}{1 - p(rated|i_k)} \geq \frac{p(rel, \neg rated|i_{k+1})}{1 - p(rated|i_{k+1})}$$
$$\Leftrightarrow p(rel|\neg rated, i_k) \geq p(rel|\neg rated, i_{k+1})$$

Which is true by description of R. Hence, swapping i_k and i_{k+1} decreases $\mathbb{E}[P@1|R]$. □

We thus get a variation of the PRP, stating we should rank items by decreasing value of $p(rel|\neg rated, i)$ rather than $p(rel|i)$. The probability $p(rel|\neg rated, i)$ corresponds to the fraction of unobserved (unrated) user tastes that are positive: the ratio of positive missing ratings. This means that the best items to be recommended are not exactly the ones that please most people, but the ones for which most unobserved preferences by the system (or undiscovered by users themselves) are positive. If we look at preference discovery as a retrieval process (prior to recommendation) in its own, $p(rel|\neg rated, i)$ represents the false negative ratio of this process. We may thus refer to this finding as the Discovery False Negative Principle (DFNP).

This principle makes natural sense in the recommendation context. An item that many people like (pure probability of relevance), but that most people have already interacted with, is of little use for recommendation, as it will be excluded from the rankings delivered to their potential "likers", and will be recommended to people who have not yet interacted with the item, but who may possibly not like it. Items with a high positive ratio in their missing ratings, in contrast, have a safe unexploited potential market –be it small or large– to make profit from.

Ratings come to be by users becoming aware of the existence of an item in the first place (by searching, browsing, advertisement, advice from a friend, random chance, etc.) and, second, by the system witnessing the encounter between the user and the item. Thus recommendation should favor items for which prior discovery has most failed, which to much extent describes the raison d'être of recommendation: complementing and filling the gaps left by other means for discovery and retrieval.

Note that we have not introduced an explicit user variable in any of the equations so far. This does not mean however the scope of our findings is restricted to non-personalized recommendation. Quite the contrary, the user variable can be assumed to be implicit in all the statements, e.g. the optimal ranking for a specific target user $u \in \mathcal{U}$ is by decreasing value of $p(rel|\neg rated, i, u)$. But since the user variable was not needed in our developments, it can also be explicitly excluded from their interpretation, and we can apply our findings in a non-personalized scope as well.

3.2 Low Discovery Recall Principle

Relevant recommendations are useful, but it is well understood in the field that relevant *and novel* ones are definitely more useful, and typically the whole purpose of recommendation [4]. Excluding rated items is a trivial realization of this principle, but unrated recommended items might still be unsurprising for the user. With this perspective in mind, we may consider accuracy metric variants that take novelty into account by simply counting as relevant only the items that the target user had not seen before.

We may for this purpose introduce an additional binary random variable $seen: \mathcal{U} \times \mathcal{I} \to \{0,1\}$ to our analysis, and consider an "undiscovered precision" metric UP such that $UP@1 = 1$ if $rel \wedge \neg seen = 1$ for the first recommendable item in the ranking. Any other relevance-oriented metric can be adapted in just the same way. Many metrics for measuring novelty have been proposed in the field [4], but this one is just direct. It is not possible to compute it with common available public datasets, but we will show an experiment where we arrange for doing so.

Now by similar steps as we followed for $P@1$, it is easy to see that the optimal ranking for $UP@1$ is by $p(rel, \neg seen|\neg rated, i) = p(\neg seen|rel, i)p(rel|i)/p(\neg rated|i)$. We get a new, even more explicit principle here: along with a high probability of relevance, items with a low prior discovery recall $p(seen|rel, i)$ are most desirable – we may refer to this as the Low Discovery Recall Principle (LDRP).

4 NON-PERSONALIZED RECOMMENDATION

Considering the principle that drives the best possible recommendation, we may wonder if we could use it to the benefit of designing the best possible recommendation algorithms, namely by seeking some approximation to $p(rel|\neg rated, i)$ and $p(rel, \neg seen|\neg rated, i)$. A proper estimation of these probabilities requires some relevance knowledge, of which a recommender system is only supplied a sample, namely, the relevance that is observed by ratings. Unfortunately using this sample is incompatible with the estimation of a probability that negates the presence of ratings as a condition.

We can however consider combinations of probabilities that may partially match the optimal ranking functions, taking ratings as an observed sample of the relevance and discovery data, in the hope that such functions may produce rankings that are, in practice, not that far from the optimal. As a simplification, we explore here non-personalized rankings, keeping the user variable away from the probabilities. Three meaningful and common non-personalized rankings can be defined in terms of ratings and relevance:

$$pop(i) = p(rated|i)$$
$$rpop(i) = p(rated, rel|i)$$
$$avg(i) = p(rel|rated, \mathrm{i})$$

The first function ranks items by their total number of ratings, commonly known as popularity in the literature [5], which is increasingly often included as a sanity check baseline in recommender systems experiments. The second function *rpop* is similar but only counts positive ratings [9]. Finally, avg is the ratio of users who have expressed a positive preference for the item, which can be read as the average rating when ratings are binary, and has been seen to perform below popularity in terms of ranking quality [5]. We explore in the next section how these non-personalized recommendations perform in relation to the optimal ranking, and to each other.

5 EXPERIMENT

To match the implicit assumptions of our theoretical analysis, we take a crowdsourced dataset that provides the opportunity to get ratings in the way users might produce through spontaneous activity, but at the same time includes further relevance and discovery knowledge that would not be obtained in the natural process.

5.1 Dataset

The dataset[1] was built using the CrowdFlower[2] platform, and includes preference judgments entered by 1,000 people for 1,000 music tracks randomly sampled from the Deezer database.[3] A judgment declares whether or not the user likes the music, after listening to a short clip of the track. Each user is assigned 100 tracks, sampled uniformly at random, in such a way that each track gets about 100 judgments, amounting to a total of around 100,000 judgments in the dataset. In addition to her taste, the user is asked whether or not she knew the music before this survey. Fig. 1a shows the user interface where the CrowdFlower workers enter their input for a music track, and Fig. 1b shows the resulting distributions of the total number of judgments, positive judgments, and prior awareness for each item.

Now we use this offline dataset to reproduce an online recommendation scenario as follows. The judgments for music that users declare having already heard before can be taken to reasonably represent ratings that users might have entered spontaneously in a system, had they come to find such items within such a system. These judgments therefore make up a reasonable representation of the input data that a recommender system is commonly supplied with. And the remaining judgments, for music that users had never heard before the survey, can be used as relevance judgments for evaluation –they apply to unrated items, the ones that are recommendable for each user. This relevance knowledge is not complete: our crowdsourced survey only covers about 10% of all items for each user. But since the user-item pairs are sampled uniformly at random, the judgments provide an unbiased estimate of the full relevance information.

a) Music judgment questionnaire b) Crowdsourced data distribution

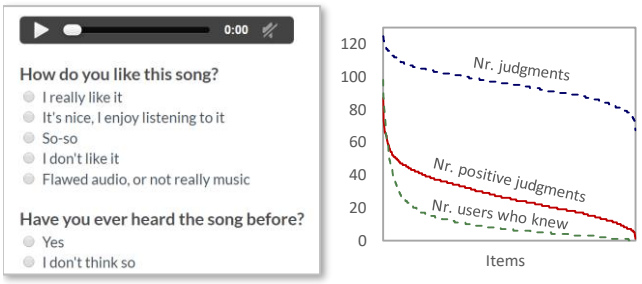

Figure 1: Music track judgment questionnaire (left) and data distribution in the obtained dataset (right). We take the top two answers to the first question in the judgment form ("how do you like this song") as indicating relevance, and the next three as non-relevance. The questionnaire does not show the song title or artist in order to get as much spontaneous and unbiased answers from users as possible. In the data distribution graph (right), the x axis for each curve is sorted by decreasing value of the y coordinate.

5.2 Standard Accuracy

To represent the design of an offline experiment, we randomly split the rating data into training and test subsets, with a ratio $\rho \in (0,1]$ of training data. The recommendation algorithms are only supplied with the training ratings, and the test data are put together with the unrated item judgments to form the set of relevance judgments for evaluation. Note that the higher the training ratio ρ, the more items shall be discarded from recommendations (because of having training ratings for more users). Thus ρ sets the transition from an offline setting with different split ratios, to an online setting experiment at $\rho = 1$ where no available input is spared for evaluation. We use this to test and observe how the experiment results may change through this transition, and see in particular how the outcome of online vs. offline experiments may agree or differ.

Fig. 2 shows the results for ρ ranging from 0.1 to 1, averaged over 100 random split repetitions for each ρ to reduce variance. For the average rating we use Dirichlet smoothing with $\mu = 1$ in the probability estimation, as it is highly sensitive to the large variance of the average value in the items with fewest ratings. Alongside the non-personalized recommendations, we evaluate the PRP and DFNP as oracle rankings that are given access to all the available relevance information. We see that for low values of ρ, the PRP and DFNP are not far from each other. However, for higher values of ρ the disagreement grows considerably due to the increasing effect of item exclusion, reaching a quite extreme point at $\rho = 1$. We see that the PRP completely fails to represent an optimal ranking at $\rho = 1$, to the point of being even substantially worse than a random recommendation. In contrast, the DFNP seems quite robust to the split ratio. A general decrease in precision with the split ratio for DFNP is natural since increasing ρ means preserving less positive relevance judgments for evaluation.

The non-personalized recommenders seem to be effective for low values of ρ, but are increasingly ineffective for higher split ratios. Popularity-based recommendation seems to follow the PRP rather than the DFNP ranking, with positive popularity $rpop$ performing slightly better than total popularity pop. In contrast, rec-

[1] The dataset is available at http://ir.ii.uam.es/cm100k.
[2] http://crowdflower.com.
[3] http://deezer.com.

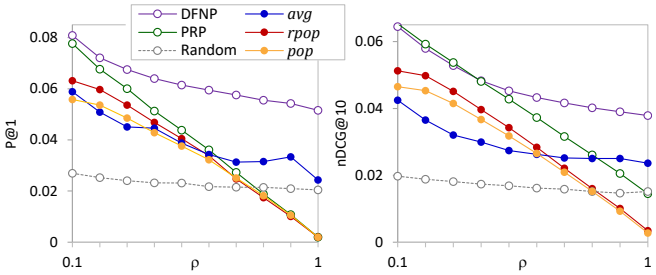

Figure 2: Experiment results. The curves show the evolution of the recommendations accuracy for different rating data split ratios by steps of 0.1.

ommendation by the average rating seems to be more robust and consistent than the popularity rankings to variations in the split ratio, and possibly a better approximation to the DFNP. It is the only ranking that stands above random recommendation for $\rho = 1$.

The poor outcome for PRP as $\rho \to 1$, and the rankings that seem to follow it, is due to the fact that the top few music tracks that most people like in the survey ("I will survive" by Gloria Gaynor, Beethoven's "Fur Elise", Mozart's "Rondo alla Turca") are known to almost everyone who was asked to judge them. As a consequence, the few users for whom the items are not excluded are mainly those who were not asked to judge them. Since we take the absence of judgment as non-relevance, this badly hurts the performance of the PRP. This may be to some extent unfair, as these items might actually please some users for whom we have no judgment. However, these users might in fact already know the items if they were asked, and again, the items would be excluded. Further research would be needed to try to elucidate what is the true situation. Be that as it may, it becomes clear that the PRP is vulnerable to the overlap between relevance and rating, and can largely diverge from an optimal ranking when these two variables strongly correlate.

5.3 Undiscovered Accuracy

Finally, we seek further insights in terms of undiscovered relevance as described in section 3.2. To be able to compute such metrics, we apply a 5-fold random split of all user judgments, taking 4/5 as training, of which only the judgments with $seen = 1$ are supplied as recommendation input; and holding out 1/5 of all judgments (including relevance and discovery information) as test data for metric computation. This makes it possible to compute regular and undiscovered versions of any accuracy metric. Fig. 3 shows the results for nDCG@10 (other metrics show a similar trend). We can see that all rankings do a terrible job at discovering useful (relevant and novel) items, except for the average rating, the only non-personalized recommendation standing above random recommendation. Note that in this setup, the undiscovered relevance in the training set is not used in the metrics computation, hence the drop of DFNP, average rating, and random ranking in standard accuracy with respect to Fig. 2, while popular items are least affected by the dropped judgments, as they had little undiscovered relevance to dispose of.

6 CONCLUSIONS

We have found that the common recommender system task, where items should not be recommended to users who have already discovered them, motivates a revision of the Probability Ranking Principle [8]. Our analysis finds simple principles for the optimal ranking in this context. We empirically confirm the divergence between these principles and the PRP in a small experiment,

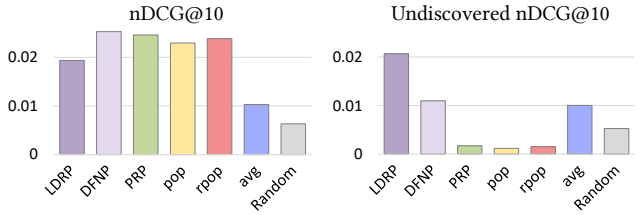

Figure 3: Standard vs. undiscovered accuracy. All pairwise differences are statistically significant (2-tailed Student's t test $p < 0.05$) except DFNP vs. PRP and *rpop* in standard nDCG (left), and PRP vs. *rpop* in undiscovered nDCG (right).

where the former show a more consistent behavior over variations in the experimental setting for recommender system evaluation. We further find that the average rating seems to display better properties than other non-personalized ranking criteria, both in terms of approximating the optimal ranking for accuracy and, yet more clearly, in delivering novel relevance.

Recent research in the field has shown that most collaborative filtering algorithms are biased towards recommending popular items [2,7]. More recently, certain algorithms have been found to be biased to the average rating instead [3], and such algorithms apparently show worse results in common experiments on public datasets. Interestingly, our present exploration raises the question whether the average rating might be a better signal than the number of ratings under certain experimental conditions, incidentally the ones that may more closely represent a live setting and true utility. This may call for a second look at the outcomes of offline experiments, under the light of further angles in the experimental design, involving e.g. the relevance judgment collection procedure, or reproducing the conditions of an online setting. Extending our analysis to personalized algorithms will likely involve the construction of larger and more dense datasets, which we envisage as future work.

ACKNOWLEDGMENTS

This work was partially supported by the national Spanish Government (grant nr. TIN2016-80630-P).

REFERENCES

[1] G. Adomavicius and A. Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17, 6 (June 2005), 734–749.

[2] A. Bellogín, P. Castells, and I. Cantador. Statistical Biases in Information Retrieval Metrics for Recommender Systems. *Inf. Ret.* 20, 6 (Dec. 2017), 606–634.

[3] R. Cañamares and P. Castells. 2017. A Probabilistic Reformulation of Memory-Based Collaborative Filtering – Implications on Popularity Biases. In *Proceedings of the 40th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR 2017), 215–224.

[4] P. Castells, N. J. Hurley, S. Vargas. 2015. Novelty and Diversity in Recommender Systems. In: *Recommender Systems Handbook, 2nd edition*, F. Ricci, L. Rokach, and B. Shapira (Eds.). Springer, New York, NY, USA, 881–918.

[5] P. Cremonesi, Y. Koren, and R. Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems* (RecSys 2010), 39–46.

[6] N. Fuhr. 2008. A probability ranking principle for interactive information retrieval. *Information Retrieval* 11, 3 (June 2008), 251–265.

[7] D. Jannach, L. Lerche, I. Kamehkhosh, and M. Jugovac. 2015. What recommenders recommend: an analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5 (Dec 2015), 427–491.

[8] S. E. Robertson. 1977. The Probability Ranking in IR. *Journal of Documentation* 33, 4 (Jan. 1977), 294–304.

[9] H. Steck. 2011. Item popularity and recommendation accuracy. In *Proc. of the 5th ACM Conference on Recommender Systems* (RecSys 2011), 125–132.

[10] M. Wechsler and P. Schäuble. 2000. The Probability Ranking Principle Revisited. *Information Retrieval* 3, 3 (Oct. 2000), 217–227.

[11] C. Zhai and J. Lafferty. 2006. A risk Minimization Framework for Information Retrieval. Information Processing & Management 42, 1 (Jan. 2006), 31–55.