Sudden Death: A New Way to Compare Recommendation Diversification (An Extended Abstract)

Derek Bridge
derek.bridge@insight-centre.org
Insight Centre for Data Analytics
University College Cork, Ireland

Mesut Kaya
mesut.kaya@insight-centre.org
Insight Centre for Data Analytics
University College Cork, Ireland

Pablo Castells
pablo.castells@uam.es
Universidad Autónonoma de Madrid
Madrid, Spain

1 INTRODUCTION

Recommendations should not only be relevant; a set of recommendations should often also be diverse.[1] A number of algorithms exist that diversify their top-N recommendation lists; a number of metrics exist that measure diversity.

But when *comparing* algorithms, especially in offline evaluations, there are at least two problems. The first problem is that the comparison may not be fair. A recommender that seeks to diversify may use an objective function that is the same as, or closely related to, the metric used to measure diversity. This metric will tend to favour this recommender. The second problem is that we are usually not interested in diversity for its own sake, although this is what most of the diversity metrics measure. The recommendations in the top-N must still be relevant to the user. Checking this means that offline experiments become multi-objective evaluations.

In this extended abstract of the paper at https://arxiv.org/abs/1908.00419, we propose a new scoring method for *comparing* algorithms. It is not specific to diversity but it is inspired by the idea that diversification is supposed to make it more likely that the user finds an item that satisfies her.

2 A CASE STUDY

We illustrate the two problems by comparing four algorithms using three diversity metrics and two relevance metrics on a single dataset. Of course, if our goal were to find the 'best' algorithm, we would use more datasets. But our purpose is simply to motivate the need for a better way of comparing algorithms.

We use the MovieLens 1 Million dataset. The algorithms we compare are MMR, xQuAD and SPAD. They greedily re-rank the recommendations made by a baseline recommender, for which we use Matrix Factorization (MF). MMR, xQuAD and SPAD measure diversity differently. MMR uses distance to the other recommendations, where distance is typically defined on item features such as movie genres [1]. xQuAD and SPAD are intent-aware diversification methods. They try to achieve coverage of the different interests ('aspects') of the user, as revealed by the user's profile. In xQuAD, the aspects are given by item features [6]. In SPAD, aspects are subprofiles, which are mined from the items that the user likes [5].

2.1 Measuring diversity

Intra-List Diversity (ILD) is a popular diversity metric; it computes the average pairwise distances of items in a recommendation set [7]. ILD may favour an algorithm like MMR [1], since ILD is close to what MMR optimizes. This is confirmed in Figure 1a.

Algorithm 1 Sudden Death

```
for $a \in Algs$ do $\mathrm{SD}(a) \leftarrow 0$
for $u \in U$ do
    for $a \in Algs$ do $h_a \leftarrow 0$
    $i \leftarrow 1$
    while $i < N$ do
        for $a \in Algs$ do $h_a \leftarrow 1$ if $\mathrm{hit}(i, u, a, N) = true$
        if $h_a = 1$ for any $a \in Algs$ then
            break
        $i \leftarrow i + 1$
    for $a \in Algs$ do $\mathrm{SD}(a) \leftarrow \mathrm{SD}(a) + h_a$
return $\mathrm{SD}(a)/|U|$ for all $a \in Algs$
```

α-nDCG is a newer diversity metric [2]. It measures coverage and relevance of aspects. We will refer to it as α-nDCG$^{\mathcal{F}}$, because aspects in this case are defined in terms of item features. Clearly, this metric favours xQuAD; and this is confirmed in Figure 1b.

In [4], we modified α-nDCG to produce a variant that we will denote by α-nDCG$^{\mathcal{S}}$, which defines aspects in terms of subprofiles. This metric favours SPAD, as confirmed by Figure 1c.

Researchers who want to make a fairer comparison end up using multiple measures and then trying, usually informally, to identify the algorithm that is most robust across those multiple measures.

2.2 Measuring the relevance-diversity trade-off

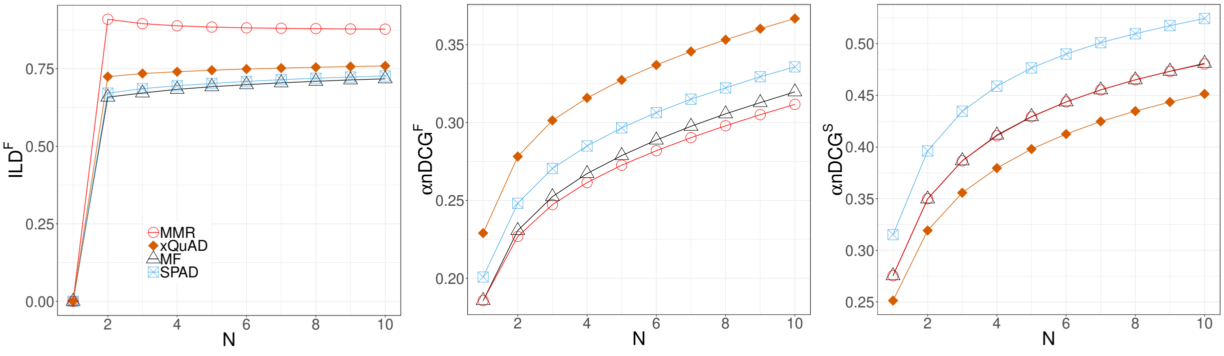
As we said before, we do not desire diversity in a top-N for its own sake. We care about relevance too. In Figures 2a and 2b, we show the precision and mean reciprocal ranks (MRR) of the different recommenders. But now to identify the best algorithm requires researchers to look jointly at relevance graphs and diversity graphs. There is a problem in deciding how to balance performance in one kind of metric against performance in another. Researchers could define metrics that combine measures of relevance and diversity but this still leaves the problem of knowing how much weight this combined metric should give to each component.

3 SUDDEN DEATH SCORE

The Sudden Death score offers a new way to compare recommenders, whether they diversify or not: user by user, it rewards the algorithms that score hits earliest in the top-N.

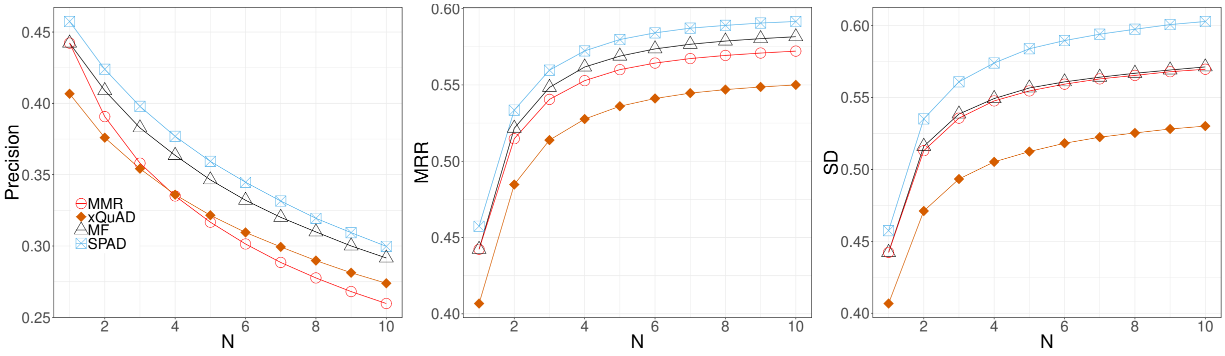
Let U be the set of users. Let rel_u be test set items that are relevant to user $u \in U$. Let *Algs* be the set of recommender algorithms to be compared. Let $\mathrm{RL}_{u,a,N}$ be the ordered list of the top-N recommendations that algorithm $a \in Algs$ makes to user $u \in U$. For a given user u, an algorithm a scores a hit at position i if any of

[1] This abstract emanates from research supported by a grant from Science Foundation Ireland (Grant Number SFI/12/RC/2289), which is co-funded under the European Regional Development Fund.



(a) ILD **(b)** α-nDCG$^{\mathcal{F}}$ **(c)** α-nDCG$^{\mathcal{S}}$

Figure 1: Different diversity metrics for different values of N.



(a) Precision **(b) MRR** **(c)** SD

Figure 2: Precision, MRR and Sudden Death score for different values of N.

the first i members of $\mathrm{RL}_{u,a,N}$ are in rel_u: $\mathrm{hit}(i, u, a, N)$ is true iff $\mathrm{RL}_{u,a,N}[: i] \cap \mathrm{rel}_u$ is non-empty. For user u, Algorithm 1 assigns recommender a a score of 1 iff no other algorithm in *Algs* has an earlier hit. It averages this across all users. The analogy with sudden death tie-breakers in sports such as badminton and volleyball should be clear: play stops in the tie-break period of the game as soon as one team is ahead. Figure 2c plots Sudden Death for different values for N for MF, MMR, xQuAD and SPAD.

There are similarities between the Sudden Death score, which is for offline evaluation, and the framework for online evaluation described in, e.g [3]. In their framework, users choose recommendations from either a single recommendation list that has been created by interleaving the results of multiple recommender systems or from multiple recommendation lists. No actual metrics are defined in [3] but one option is to reward a system if it places the item chosen by the user earlier in its recommendation list.

The Sudden Death score tries to capture the idea that diversification should make it more likely that the user finds an item that satisfies her. One concern might be that it seems very similar to relevance measures such as precision and MRR, which also reward algorithms for hits and, in some cases, for earlier hits. But Figure 2 shows that they are different: although here SPAD is always the best, they rank MF, MMR and xQuAD differently depending on N.

But there is a more fundamental difference too. Precision and MRR are performance estimates. We compute them to get an estimate of how well a model will perform on future unseen data. The Sudden Death score is not a performance estimate. Its purpose is only for *comparing* systems. It follows that the Sudden Death score does not replace precision and MRR; it supplements them.

REFERENCES

[1] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *21st ACM SIGIR Intl. Conference on Research & Development in Information Retrieval.* 335–336.

[2] Charles L A Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *31st ACM SIGIR Intl. Conf. on Research and Development in Information Retrieval.* 659–666.

[3] Conor Hayes and Pádraig Cunningham. 2002. *An on-line evaluation framework for recommender systems.* Technical Report. Trinity College Dublin, Department of Computer Science.

[4] Mesut Kaya and Dereek Bridge. 2019. A Comparison of Calibrated and Intent-Aware Recommendations. In *13th ACM Conf. on Recommender Systems.*

[5] Mesut Kaya and Derek Bridge. 2019. Subprofile-aware diversification of recommendations. *User Modeling and User-Adapted Interaction* (2019).

[6] Saúl Vargas Sandoval. 2015. *Novelty and Diversity Evaluation and Enhancement in Recommender Systems.* Ph.D. Dissertation. Universidad Autónoma de Madrid.

[7] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *14th Intl. Conf. on World Wide Web.* 22–32.