Demonstrating the Equivalence of List Based and Aggregate Metrics to Measure the Diversity of Recommendations (Student Abstract)

Maurizio Ferrari Dacrema
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy
maurizio.ferrari@polimi.it

Abstract

The evaluation of recommender systems is frequently focused on accuracy metrics, but this is only part of the picture. The diversity of recommendations is another important dimension that has received renewed interest in recent years. It is known that *accuracy* and *diversity* can be conflicting goals and finding appropriate ways to combine them is still an open research question. Several ways have been proposed to measure the diversity of recommendations and to include its optimization in the loss function used to train the model. Methods optimizing list based diversity suffer from two drawbacks: the high computational cost of the loss function and the lack of an efficient way to optimize them. In this paper we show the equivalence of the list based diversity metrics Hamming and Mean Inter-List diversity to the aggregate diversity metric measured with the Herfindahl index, providing a formulation that allows to compute and optimize them easily.

Introduction

The importance of providing the user with diverse recommendations has been known for several years, but enhancing diversity often adversely affects the model accuracy. Addressing this known trade-off is still an open research question that has driven the creation of several approaches to jointly optimize both (Kaminskas and Bridge 2017). Commonly used diversity metrics can be classified in three different categories: *individual*, *aggregate* and *list based* diversity. While individual diversity only measures what is perceived by the user and is computed on each separate recommendation list, aggregate diversity considers the system as a whole and is measured taking into account the recommendations provided to all users. Aggregate diversity metrics can be computed and optimized easily. An example of them is the Item coverage, which represents the quota of items that have been recommended at least once. Note that individual and aggregate diversity can behave very differently. A higher aggregate diversity is often a desirable property for a recommender system, since it will more likely encourage the users to explore less popular items (i.e., long tail items) and may prove beneficial in domains prone to high popularity bias which could result in poor catalogue coverage. Aggregate diversity is also useful to gain a system-wide overview, which is important when the recommender is part of a business model (Brynjolfsson, Hu, and Simester 2011). List based diversity metrics can be computed from the recommendation lists received by all users. Although often used in the literature, these metrics suffer from several drawbacks: they have a very high computation cost, their formulation is not easy to optimize directly and their relation to other diversity metrics is not known analytically. These issues constitute significant obstacles for the development of new algorithms aiming to improve the state of the art. In this paper we consider the *mean inter-list* diversity (MIL), proposed by (Zhou et al. 2010), which represents the average number of recommendations any pair of users has in common. We first show that MIL is equivalent to the average Hamming distance in a set of strings having all equal length. Secondly, we demonstrate that both can be computed based solely on the number of times each item appears in any recommendation list (or string), due to this both metrics are aggregate diversity metrics equivalent to the *Herfindahl index* and can be computed and optimized easily. Lastly, we show how a reranking strategy, previously demonstrated only at an empirical level, optimizes MIL diversity.

Diversity Metrics

Among the most used aggregate diversity metrics are Item Coverage, Shannon Entropy, Gini Index and Herfindahl index (Zhou et al. 2010; Paudel et al. 2017), all are functions of the global number of times each item has been recommended but measure different behaviors, some being quadratic, logarithmic or only accounting for items once.

The following notation will be adopted in the paper. The item set is I, the user set U and the respective cardinality $|I|$, $|U|$. The length of the recommendation list, i.e., the cut-off, is c, while $rec(i)$ represents the number of times item i has been recommended across all users. The total number of recommendations is $rec_t = \sum_{i \in I} rec(i) = c \cdot |U|$. Of particular interest for this paper are the following metrics:

Herfindahl index: Is an aggregate diversity metrics computed with the square of the number of times each item has been recommended. It is easy to compute and optimize.

$$\text{Herfindahl} = 1 - \frac{1}{rec_t^2} \sum_{i \in I} rec(i)^2$$

Mean Inter-List diversity (MIL): This diversity (Zhou et al. 2010)[1] considers the uniqueness of different user's recommendation lists and has a value between 0 and 1. It is easily interpretable, the less likely any two users have been recommended the same items, hence the more diverse the recommendations are, the closer MIL will be to 1. MIL is computed as an average over all inter-list distances, excluding the diagonal, where *ua* and *ub* are two users and $q(ua, ub)$ is the number of common items in their recommendation lists.

$$MIL = \frac{1}{|U|^2 - |U|} \sum_{\substack{ua,ub \in U \\ ua \neq ub}} 1 - \frac{q(ua, ub)}{c}$$

This formulation has two crucial issues. First, it is not clear how it relates to other aggregate diversity metrics and it is difficult to optimize. Second, computing MIL requires to compute function $q(ua, ub)$ for all couples of users, which is quadratic in their number and very computationally expensive for all but the smallest datasets. The computational complexity is actually the same of building a user-based nearest neighbor model, known for its low scalability.

Hamming diversity: Is defined on the user recommendation lists L, represented as a one-hot encoding L_H of $|I|$ elements. The Hamming distance is the number of positions in which the two lists are different. Since the Hamming distance can be computed from $q(ua, ub)$ as $H(ua, ub) = |I| - q(ua, ub)$ Hamming and MIL diversity are equivalent.

Proof of Equivalence

We now demonstrate that MIL and Hamming diversity metrics are aggregate diversity metrics and can be computed based solely on the global item distribution. Being an arithmetic mean both metrics are defined for recommendation lists of a fixed length c and would otherwise produce erroneous results. The computationally expensive part of MIL is the summation of common items in all recommendation lists. We can isolate this component as q_g. We then decompose function $q(ua, ub)$ as a summation of other functions $q_i(ua, ub)$, each associated to a specific item i, that will have value 1 if both users have been recommended item i, 0 otherwise. Finally, we swap the two summations.

$$MIL = 1 - \frac{1}{|U|^2 - |U|} \frac{q_g}{c}$$

$$q_g = \sum_{i \in I} \sum_{\substack{ua,ub \in U \\ ua \neq ub}} q_i(ua, ub) = -|U|c + \sum_{i \in I} rec(i)^2$$

This formulation allows to represent q_g in terms of a combinatorial problem that can be easily solved. The summation of $q_i(ua, ub)$ over all pairs of users that received item i in their recommendation list is equal to the number of non-ordered pairs that can be defined from such set. Since the set contains $rec(i)$ users, the number of non-ordered pairs it allows is $rec(i) \cdot (rec(i) - 1)$. The global common item count q_g can therefore be represented directly in terms of the global item distribution. This result indicates that MIL diversity, Hamming diversity and Herfindahl index are equivalent, being linear functions of the same quadratic summation of $rec(i)$. Hence, all three diversity metrics measure aggregate diversity and do not depend on the specific user recommendation lists but only on the final item occurrence distribution, which is easily available and allows to compute them in negligible time. Due to their quadratic nature, they will have low sensitivity to items having low number of occurrences and high sensitivity to items being recommended often. Another important consequence of what previously demonstrated is that it is possible to control or optimize the MIL diversity of a recommender system by altering the average probability each item will appear in the recommendation lists. This can be achieved during the training of the model including it in the loss function or, for example, via a reranking step.

[1]Note that MIL was called *Personalization*. We will not use this name because it can be maximized by random recommendations.

Diversity Enhancing Reranking

This new formulation of MIL diversity is also useful to explain previously published experimental results for algorithms that implemented reranking steps to improve the diversity of recommendations. In particular, (Paudel et al. 2017) propose RP$^3\beta$, a graph-based collaborative recommendation algorithm that applies a reranking step, dividing the score of the item as computed by a previous algorithm, P$^3\alpha$, by their popularity, in order to penalize very popular items. Both algorithms provide very competitive recommendation quality even against state of the art neural models (Ferrari Dacrema, Cremonesi, and Jannach 2019). This reranking step was proposed with an intuitive justification, we are now able to explain its connection with the diversity metrics. Since P$^3\alpha$ is very influenced by the item popularity, this value constitutes a good approximation of the $rec(i)$ function, therefore RP$^3\beta$ is optimizing an approximation of MIL diversity. A similar reranking approach could be also adapted and applied to other recommendation models.

Acknowledgements

I am grateful to prof. Paolo Cremonesi for his support.

References

Brynjolfsson, E.; Hu, Y.; and Simester, D. 2011. Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales. *Management Science* 57(8): 1373–1386.

Ferrari Dacrema, M.; Cremonesi, P.; and Jannach, D. 2019. Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches. *RecSys* 101–109.

Kaminskas, M.; and Bridge, D. 2017. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM TiiS* 7(1): 2.

Paudel, B.; Christoffel, F.; Newell, C.; and Bernstein, A. 2017. Updatable, Accurate, Diverse, and Scalable Recommendations for Interactive Applications. *ACM TiiS* 7(1): 1.

Zhou, T.; Kuscsik, Z.; Liu, J.-G.; Medo, M.; Wakeling, J. R.; and Zhang, Y.-C. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *National Academy of Sciences* 107(10): 4511–4515.