Post-hoc Selection of Pareto-Optimal Solutions in Search and Recommendation

Vincenzo Paparella
Politecnico di Bari
Bari, Italy
vincenzo.paparella@poliba.it

Vito Walter Anelli
Politecnico di Bari
Bari, Italy
vitowalter.anelli@poliba.it

Franco Maria Nardini
ISTI-CNR
Pisa, Italy
francomaria.nardini@isti.cnr.it

Raffaele Perego
ISTI-CNR
Pisa, Italy
raffaele.perego@isti.cnr.it

Tommaso Di Noia
Politecnico di Bari
Bari, Italy
tommaso.dinoia@poliba.it

ABSTRACT

Information Retrieval (IR) and Recommender Systems (RSs) tasks are moving from computing a ranking of final results based on a single metric to multi-objective problems. Solving these problems leads to a set of Pareto-optimal solutions, known as Pareto frontier, in which no objective can be further improved without hurting the others. In principle, all the points on the Pareto frontier are potential candidates to represent the best model selected with respect to the combination of two, or more, metrics. To our knowledge, there are no well-recognized strategies to decide which point should be selected on the frontier in IR and RSs. In this paper, we propose a novel, post-hoc, theoretically-justified technique, named "Population Distance from Utopia" (PDU), to identify and select the one-best Pareto-optimal solution. PDU considers fine-grained utopia points, and measures how far each point is from its utopia point, allowing to select solutions tailored to user preferences, a novel feature we call "calibration". We compare PDU against state-of-the-art strategies through extensive experiments on tasks from both IR and RS, showing that PDU combined with calibration notably impacts the solution selection.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Retrieval models and ranking**.

KEYWORDS

Pareto optimality, Information Retrieval, Recommender Systems

ACM Reference Format:
Vincenzo Paparella, Vito Walter Anelli, Franco Maria Nardini, Raffaele Perego, and Tommaso Di Noia. 2023. Post-hoc Selection of Pareto-Optimal Solutions in Search and Recommendation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3615010

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10.
https://doi.org/10.1145/3583780.3615010

1 INTRODUCTION

Many tasks in Information Retrieval (IR) and Recommender Systems (RSs) involve the optimization of multiple objective functions. As an example, consider the IR task of *diversifying search results* where, given a user query, we require the IR system to return a list of results that are both *relevant* for the user and *diverse* concerning the possible "facets" of the query [40]. Addressing this task asks for designing a two-objective ranking function comprehensively maximizing both the relevance and the diversity of the result list. The same considerations can be made in RSs. Despite the accuracy of recommendation being considered the gold measure to assess the quality of suggestions, over the last years, RSs have been required to meet other *beyond-accuracy* metrics to avoid obvious [46] and unfair [51] recommendations. Therefore, the choice of a recommendation model and its setting entail several criteria leading to a trade-off among them, resulting in a non-trivial challenge.

Multi-Objective Optimization (MOO) recently attracted several interesting IR and RS contributions [15, 41, 51]. MOO deals with *Pareto optimality*, i.e., the identification of solutions where no objective can be further improved without damaging the others. Pareto-optimal solutions are in turn collected in the so-called *Pareto Frontier*, a set of (possibly infinite) non-dominated solutions.

Existing approaches for MOO can be classified into two categories: i) *heuristic search* and, ii) *scalarization*. In the first category, multi-objective evolutionary algorithms are used to ensure that the emerging solutions are not dominated by each other, even if they can still be dominated by Pareto-optimal solutions not visited by the algorithm [6, 39]. In the second category, scalarization methods aggregate multiple objectives into one objective, possibly guaranteeing Pareto optimality. Scalarization approaches can exploit *model aggregation* techniques combining the output of different models trained on the single objectives. Alternatively, *label aggregation* techniques combine the labels of the training samples a priori, and the optimization is performed using the aggregated labels. Aggregation techniques may involve the setting of the importance or priority of the different objectives by weighting each objective through a scalar function (e.g., Linear Scalarization [31], Weighted Chebyshev [27]). Conversely, some techniques work by constraining the objectives of the problem, e.g., ϵ–Constraint [17] leading to a unique non-dominated solution.

Pareto optimality is commonly achieved by many different Pareto-optimal solutions. However, IR and RS MOO tasks generally require

identifying a single Pareto-optimal solution to be deployed in the system. To the best of our knowledge, no strategies specifically tailored to IR and RS tasks have been previously proposed [51]. The state-of-the-art techniques from MOO theory are in fact aimed at identifying a set of Pareto-optimal solutions, without addressing the problem of *post-hoc* choosing one among the—possibly many—solutions identified for the IR and RS tasks. Indeed, many works in the IR and RS literature, although exploiting the techniques discussed above, do not either: i) consider the problem of selecting a single best solution to the multi-objective problem or, ii), discuss the criteria adopted to select a single Pareto-optimal solution [53].

In this paper, we fill this gap by introducing "Population Distance from Utopia" (PDU), a novel post-hoc flexible strategy for selecting **one**—**best**—Pareto-optimal solution among the ones lying in the Pareto frontier for IR and RS tasks. PDU relies on the observation that the Pareto-optimal point coordinates are an aggregation—usually the mean—of the model performance for each sample, i.e., queries in IR and users in RS, on multiple objectives. PDU exploits the notion of "Utopia point" as the ideal optimization target. Differently from the methods from MOO theory, which are devised to solely consider the mean performance values when selecting a single Pareto-optimal solution, PDU is designed to set a utopia point for each sample of the dataset. This feature allows choosing a solution not only based on the "global" performance achieved by the IR/RS model, but also in a more fine-grained resolution that now considers multiple quality criteria that are expressed on a sample level. We call this feature "calibrated" selection. In detail, the novel contributions of this paper are:

- We formally introduce PDU as a novel technique that allows one to select, in a principled way, the best Pareto-optimal solution previously identified by a state-of-the-art MOO technique.
- We provide a thorough comparison of PDU against state-of-the-art selection strategies. The analysis shows that PDU is the only selection method that allows identifying a "calibrated" solution, i.e., based on ideal targets expressed on a sample level.
- We experimentally compare PDU against state-of-the-art strategies on well-known IR and RS tasks by exploiting public data. The results show that, unlike other methods, PDU can identify Pareto-optimal solutions regardless of their position on the frontier. Moreover, PDU calibration can lead to the selection of significantly different trade-offs.
- We release a GitHub repository[1] for our implementation of PDU and the state-of-the-art competitors as well as the data used in the experiments to allow a full reproducibility of the results.

2 MULTI-OBJECTIVE OPTIMIZATION

A Multi-Objective Optimization Problem (MOOP) [31] is defined as:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \boldsymbol{f}(\mathbf{x}) = \{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})\} \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{1}$$

The vector $\mathbf{x} \in \mathbb{R}^n$ is formed by n independent variables called *decison variables*. The set $\mathcal{X} \subseteq \mathbb{R}^n$, generally known as *feasible set*, is defined by a set of equality and inequality constraints such as $\{\mathbf{x} \,|\, g_j(\mathbf{x}) \leq 0, j = 1, 2, \ldots, l; \vee\, h_i(\mathbf{x}) = 0, i = 1, 2, \ldots, e\}$. The vector of functions $\boldsymbol{f}(\cdot)$ is composed by $k \geq 2$ scalar *objective functions* $f_i : \mathcal{X} \to \mathbb{R}$ with $i = 1, \ldots, k$. In multi-objective optimization, the space $\mathbb{R}^k$ is known as *objective function space*.

[1] https://github.com/sisinflab/Selection-Pareto-Optimal-Solutions-IR-RS

Pareto Optimality. In a MOOP, since typically there is no single global solution, it is impossible to determine a set of points that all fit a predetermined definition for an optimum. Hence, it is usually adopted the concept of *Pareto optimality* which leverages on the *Pareto dominance* relation [47]. A vector $\mathbf{x}^\star$ Pareto-dominates vector $\mathbf{x}$, denoted by $\mathbf{x}^\star \prec \mathbf{x}$, if and only if $\exists\, j \in \{1, \ldots, k\} \,|\, f_j(\mathbf{x}^\star) < f_j(\mathbf{x})$ and $f_i(\mathbf{x}^\star) \leq f_i(\mathbf{x})\ \forall i \in \{1, \ldots, j-1, j+1, \ldots, k\}$. We also write that, a solution $\mathbf{x}^\star \in \mathcal{X}$ is Pareto optimal if there does not exist another solution $\mathbf{x} \in \mathcal{X}$ such that $\mathbf{f}(\mathbf{x}) \prec \mathbf{f}(\mathbf{x}^\star)$. In other words, a point is Pareto optimal if there is no other point that improves at least one objective function without hurting another one. Then, solving the problem in Equation (1) means finding the solutions $\mathbf{x} \in \mathcal{X}$ such that their images $\mathbf{f}(\mathbf{x})$ are not Pareto-dominated by any other vector in the feasible set. The set of non-Pareto-dominated solutions $P^\star \subseteq \mathcal{X}$ is called Pareto-optimal set in the feasible set, that is formally defined as $P^\star := \{\mathbf{x}^\star \in \mathcal{X} | \neg \exists\, \mathbf{x} \in \mathcal{X} \text{ s.t. } \mathbf{x} \prec \mathbf{x}^\star\}$. The image of the Pareto-optimal set $P^\star$ in the objective function space is called the Pareto frontier, i.e., $PF^\star := \{\boldsymbol{f}(\mathbf{x}^\star) \,|\, \mathbf{x}^\star \in P^\star\}$.

Utopia and Nadir Points. Once a solution $P^\star$ for the problem in Equation (1) is obtained, the decision-making process requires the selection of a single optimal solution from the Pareto frontier. Generally, the *utopia point* helps to implement this process [31]. A point $\boldsymbol{f}^\diamond \in \mathbb{R}^k$ is a utopia point if and only if $f_i^\diamond = \min_{\mathbf{x}} f_i(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}\ \forall i \in \{1, 2, \ldots, k\}$. Generally, the utopia point is the *ideal* point in $\mathbb{R}^k$ that is unattainable. Hence, a common approach consists in reaching the *closest* solution to the utopia point as the best one, where, in most of the cases, the term *closest* refers to the solution which minimizes the Euclidean distance to the utopia point. However, it is not necessary to restrict closeness to the case of a Euclidean norm [31].

Along with the utopia point, the *nadir point* also helps select a solution from the Pareto frontier. Dually to the utopia point, the nadir point represents the point in the objective function space having the worst possible values for each objective. A point $\boldsymbol{f}^\triangle \in \mathbb{R}^k$ is a nadir point if and only if $f_i^\triangle = \max_{\mathbf{x}} f_i(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{X}\ \forall i \in \{1, 2, \ldots, k\}$. Compared to the utopia point, determining the nadir point can be challenging, even for simple problems [26].

3 BACKGROUND

3.1 Selection Strategies

The Pareto frontier consists of a set of equally optimal solutions. Some methods to select a single Pareto-optimal solution assume the existence of a decision maker [25]. These methods are known as *Multi-Criteria Decision Making* (MCDM) strategies, where a decision-maker has knowledge of the preferences (hierarchy) among the objectives. However, decision-makers do not always know how to weigh the different objectives [5]. Moreover, in some cases, the complexity of the problem makes it difficult for a human decision-maker to evaluate and compare different options comprehensively. Conversely, mathematical methods can provide consistent, objective, and impartial decision-making approaches. In this work, we focus and outline mathematical strategies for selecting a solution from the Pareto frontier, i.e., strategies applicable in the absence of "a priori knowledge" that can feed an MCDM strategy.

3.1.1 Knee Point. The *Knee Point* [5] strategy aims to identify a knee of the Pareto frontier. The rationale is that solutions different from the knee point would exhibit limited improvement for one objective and a substantial deterioration for the others. As described by Branke et al. [5], these strategies were born as a variation of multi-objective evolutionary algorithms to find the knee regions on the Pareto frontier. Consequently, when other algorithms compute the Pareto Frontier, the extracted knee region may not have a knee-featured shape, thus making this strategy less convenient. Several methods to identify the knee point are proposed in the literature, mainly differing for the number of objectives.

Angle-based method (A-KP). When dealing with two objectives, the reflex angle between the slopes of the two vectors through a point $B = (x_i, y_i)$ and its two neighbors, i.e., $A = (x_{i-1}, y_{i-1})$ and $C = (x_{i+1}, y_{i+1})$, on the Pareto Frontier can be considered as an efficient indication of whether the point can be classified as a knee [5]. *The Pareto-optimal point having the maximum reflex angle computed from its neighbors is considered the knee* [12]. If no neighbor to the left (right) is found, a vertical (horizontal) line is used to calculate the angle. Even though this method is efficient in a two-dimensional scenario, it becomes impractical for more than two objectives, especially for the choice of neighbors.

Utility-based method (U-KP). A valid alternative to overcome the limitation of the angle-based method is adopting a marginal utility function. Let us consider a set of n objective functions $f(\cdot)$ and m sets of n uniformly distributed weights $\mathbf{w}$, with $w_i \in [0, 1]$ such that $\sum_i w_i = 1$ [5]. The resulting utility function is then $U(\mathbf{x}, \mathbf{w}) = \sum_i w_i \cdot f_i(x)$. The Pareto-optimal solution having the minimum utility value for most weight configurations is the knee point.

3.1.2 Hypervolume. The *Hypervolume* [54] strategy was first introduced to compare the quality of different Pareto frontiers [14]. However, by computing the hypervolume of each solution on the Pareto frontier, this strategy can be straightforwardly exploited to select the best solution from the set [53]. Given a Pareto-optimal solution $\mathbf{x}^\star \in \mathbb{R}^k$, a reference point $\mathbf{r} \in \mathbb{R}^k$, and the Lebesgue measure λ, the hypervolume $\mathcal{HV}$ of $\mathbf{x}^\star$ with respect to $\mathbf{r}$ is:

$$\mathcal{HV} = \lambda(\{\mathbf{x} \in \mathbb{R}^k \mid \mathbf{x}^\star \prec \mathbf{x} \prec \mathbf{r}\}). \tag{2}$$

The $\mathcal{HV}$ value is the volume of the hypercube determined by the solution $\mathbf{x}^\star$ and the reference point $\mathbf{r}$. *The Pareto-optimal point having the maximum hypervolume is the selected one.*

3.1.3 Other Techniques. Other simpler techniques that have been used for selecting a solution from the Pareto frontier are the *Euclidean Distance* and the *Weighted Mean* [34, 50]. The Euclidean Distance (*ED*) is computed between each solution on the Pareto frontier and the utopia point: $ED(\mathbf{x}^\star) = |\boldsymbol{f}(\mathbf{x}^\star) - \boldsymbol{f}^\diamond|$. *The Pareto-optimal point having the minimum Euclidean distance is the selected solution.* Instead, the Weighted Mean (*WM*) requires setting the importance of each objective through a set of weights. *Among all the Pareto-optimal points, the point maximizing the weighted mean corresponds to the selected solution.*

3.2 Related Works on MOO for IR and RS

Previous works investigate the introduction of multiple criteria in IR systems, e.g., in web search and recommendation [10, 11, 21, 44, 45, 49], and product search [22, 29]. Carmel et al. [9] propose Stochastic Label Aggregation (SLA), a technique that perform label aggregation by randomly selecting a label per training example. In RS, Lin et al. [28] propose a scalarization based Pareto-Efficient Learning-To-Rank (PE-LTR) framework by deriving the conditions for the weighted sum weights that ensure the solution to be Pareto efficient. In the RS area, MOO techniques are routinely exploited for optimizing multiple fairness criteria beyond relevance. Ge et al. [15] propose a fairness-aware RS based on multi-objective reinforcement learning, simultaneously optimizing clickthrough rate (CTR), as a signal for relevance, and item exposure, as a signal for fairness. Moreover, Wu et al. [51] employ scalarization to optimize accuracy along with both provider and consumer fairness. Naghiaei et al. [32] also integrate fairness constraints from a consumer and producer-side into a re-ranking approach.

4 POPULATION DISTANCE FROM UTOPIA

Driven by the goal of overcoming the limitations of the other methods in a principled way for IR and RSs, we propose PDU (Population Distance from Utopia), a selection strategy taking into account the distance of the query/user metrics from the utopia point.

Our intuition starts from the observation that in a search and/or recommendation scenario, the Pareto frontier is populated by points representing aggregated results (usually, they represent the average value) on metrics referring to a set of experiments. For instance, in a RS setting, we could have a frontier representing the values of two metrics: *nDCG*, to measure the accuracy of the model, and *Intralist Diversity* (*ID*), to measure the diversity in the list of recommended items. Each point on the frontier may represent the corresponding values of *nDCG* and *ID* for a specific configuration of the hyperparameters. It is worth noticing that these values are computed as the value of the given metric averaged on all the system users. Suppose we focus instead on the point representing the single user. In that case, we may also reconsider the notion of utopia point in this more fine-grained view and adapt it to generalize with respect to the single user. The same observations hold in a search setting where we have queries instead of users. The questions leading our proposal are then: i) *What happens if we focus our analysis on the original points instead of their aggregated representation?* ii) *Can we characterize each of these fine-grained points and exploit a generalized definition of utopia point that considers even the single user/query?* We start by defining a generalized version of the utopia point.

A point $\boldsymbol{f}^\circ$ in the *objective function space* $\mathbb{R}^k$ is a ***generalized utopia point*** if and only if $f_i^\circ = h_i(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\ \forall i \in \{1, 2, \ldots, k\}$. In our definition, h_i is a function that considers the characteristics of the original data and returns a desired but unattainable utopia value for the i-th metric. For a (non-generalized) utopia point $f^\diamond$, we have $h_i = \min_{\mathbf{x}} f_i(\mathbf{x})$. Its definition can be driven both by system or dataset properties and by the choices of the system designer. For instance, in Section 5.1, we define h_2 (see Equation (14)) to quantify the users' popularity tendencies stemming from their past interactions with the items in a recommendation scenario.

Given a Pareto-optimal solution $\mathbf{x}^\star \in \mathbb{R}^k$, we can assume that it is the image of an aggregation function applied to a set of m points $\mathbf{x}_j$ in $\mathbb{R}^k$, with $j \in \{1, \ldots, m\}$. In our previous example, the points represent the values of the pairs $\langle nDCG, ID \rangle$ (with $k = 2$) for the m

users in the system. Suppose a generalized utopia point $\boldsymbol{f}_j^\circ \in \mathbb{R}^k$, with $j \in \{1, \dots, m\}$, is associated to each point $\mathbf{x}_j$.

Definition 4.1. The *Population Distance from Utopia* (PDU) is:

$$\text{PDU} = \log\left(\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, \mathbf{x}_j)^2\right), \tag{3}$$

where $e : \mathbb{R}^k \to \mathbb{R}$ is an error function that satisfies the conditions of identity, symmetry, and triangle inequality. *The Pareto-optimal point having the minimum PDU is the selected solution.* The error function $e(\cdot)$ is parametric, i.e., we can set any error or distance metric as $e(\cdot)$, like Euclidean distance or mean squared error.

Derivation. Let us consider an objective function space $\mathbb{R}^k$, where k is the number of objectives, and a dataset $\mathcal{D}$ of m samples (users/queries). For each sample, we suppose to know the best possible value of each objective. Then, we can associate each sample with a k-dimensional vector $\boldsymbol{f}_j^\circ$, with $j \in \{1, \dots, m\}$, which constitutes its generalized utopia point in the objective function space $\mathbb{R}^k$. We use $\mathbf{F} = \{\boldsymbol{f}_j^\circ \mid j \in \{1, \dots, m\}\}$ to denote the set of all the generalized utopia points referring to the m samples. Let us now consider a model η that returns k objectives performance values for each sample in $\mathcal{D}$. As before, each sample corresponds to a k-dimensional vector $\mathbf{x}_j$, with $j \in \{1, \dots, m\}$, which represents the model performance for that sample in $\mathbb{R}^k$. We denote $\mathcal{P} = \{\mathbf{x}_j \mid j \in \{1, \dots, m\}\}$. Thus, each sample j is represented by $\boldsymbol{f}_j^\circ$ and $\mathbf{x}_j$ in the objective function space: the closer the points, the better the model η performs. Let us introduce an error function $e : \mathbb{R}^k \to \mathbb{R}$ satisfying the conditions of identity, symmetry, and triangle inequality. The error of the model η on the j-th sample is $e(\boldsymbol{f}_j^\circ, \mathbf{x}_j)$. By supposing the error term follows the IID property, it has a Gaussian distribution with mean $\mu = 0$ and variance σ^2, i.e., $e(\boldsymbol{f}_j^\circ, \mathbf{x}_j) \sim \mathcal{N}(0, \sigma^2)$, whose probability density function is:

$$p(e(\boldsymbol{f}_j^\circ, \mathbf{x}_j)) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{e(\boldsymbol{f}_j^\circ, \mathbf{x}_j)^2}{2\sigma^2}\right). \tag{4}$$

We can note that if $\boldsymbol{f}_j^\circ$ and $\mathbf{x}_j$ are close, the exponent part of Equation (4) tends to 1, and the probability increases while tending to 0 when the two points are far apart and the probability decreases.

Then, we compute the error probability density function of the error for the entire dataset $\mathcal{D}$. We observe that the model η has some parameters $\boldsymbol{\Theta}$. Hence, $\mathcal{P}$ can be expressed as a function g of the parameters $\boldsymbol{\Theta}$: $\mathcal{P} = g(\boldsymbol{\Theta})$. Then, a vector $\mathbf{x}_j \in \mathcal{P}$ can be rewritten as $\mathbf{x}_j = g(\boldsymbol{\Theta})_j$. By assuming the samples to be independent, we obtain the following expression for the likelihood function:

$$p(e(\mathbf{F}, g(\boldsymbol{\Theta}))) = \prod_{j=1}^{m} p(e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)). \tag{5}$$

Since $\boldsymbol{f}_j^\circ$ is the (generally unattainable) output we desire to have, we are interested in finding the parameters $\boldsymbol{\Theta}$ for the model η such that the likelihood function $p(e(\mathbf{F}, g(\boldsymbol{\Theta})))$ is the highest. As the logarithmic function is increasing monotone, it does not modify the maximum positions. Hence, we can compute the log-likelihood instead of the likelihood to simplify calculations:

$$\log p(e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta}))) = \log \prod_{j=1}^{m} p(e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)) \tag{6}$$

$$= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{2\sigma^2} \sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2. \tag{7}$$

At this point, we explicit the variance term σ^2. Since we have supposed that the error term $e(\boldsymbol{f}_j^\circ, \mathbf{x}_j)$ has a Gaussian distribution with $\mu = 0$, the variance σ^2 is defined as $\frac{\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2}{m}$. By introducing this term in Equation (7), we obtain that the log-likelihood is:

$$\begin{aligned} \log p(e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta}))) = m \log \frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{m}\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2}} \\ - \frac{1}{\frac{2}{m}\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2} \sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2 \end{aligned} \tag{8}$$

$$= -m \log(\sqrt{2\pi}) + m \log m - \frac{1}{2} \log\left(\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2\right) - \frac{m}{2}. \tag{9}$$

By supposing to train the model η on the same dataset $\mathcal{D}$ with several configurations of $\boldsymbol{\Theta}$, the terms depending on the dataset size m and the constant $1/2$ in Equation (9) can be removed as they are constant when choosing the highest log-likelihood. Hence, the only variable quantity among the different log-likelihoods is:

$$-\log\left(\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2\right). \tag{10}$$

Therefore, we are looking for the model η with parameters $\boldsymbol{\Theta}$ having the maximum value of the term in Equation (10):

$$\max\left[-\log\left(\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2\right)\right]. \tag{11}$$

Finally, this remainder term can be easily rewritten with a positive sign as long as we choose the configuration of $\boldsymbol{\Theta}$ for the model η having the minimum value for this quantity:

$$\min\left[\log\left(\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, g(\boldsymbol{\Theta})_j)^2\right)\right] = \min\left[\log\left(\sum_{j=1}^{m} e(\boldsymbol{f}_j^\circ, \mathbf{x}_j)^2\right)\right]. \tag{12}$$

□

4.1 Calibrated PDU

PDU allows setting a generalized utopia point for each sample of the dataset, i.e., queries and users in an IR or RS scenario, respectively. This feature allows choosing a solution not only based on the "global" performance achieved by the IR/RS model, but also in a more fine-grained resolution that now considers multiple quality criteria expressed on a sample level. We call such feature ***calibration*** since it can be usefully exploited in specific scenarios, e.g., personalization in RS, where it is possible to define generalized utopia points according to individual users' preferences. These generalized utopia points can be fixed apriori, e.g., they can be identified by the system designer or computed through functions that numerically quantify the users' tendencies, similarly to what has been done in previous

works regarding *calibrated recommendations* [20, 35, 42]. We refer to this feature as *Calibrated*-PDU (C-PDU).

4.2 Feature Comparison

In Section 3.1, we have presented the most-used techniques to choose a single best solution belonging to a Pareto frontier. However, as also stated by Wu et al. [51], there is no consensus on the strategy to solve this task in the IR and RS communities. Not surprisingly, all methods have some advantages and limitations, leading to a lack of an ideal strategy [26]. Hence, a comparison of the features provided by PDU and state-of-the-art techniques is needed. Specifically, we identify some desirable features the techniques should have. Table 1 discusses the main properties of PDU and other state-of-the-art techniques. First, **the strategy should be suitable even when dealing with more than two objectives**. In this regard, the angle-based knee point is the only ineffective method. Second, **the strategy should not need any additional knowledge**. Most techniques require additional problem information, i.e., the reference point ($\mathcal{HV}$), the (generalized) utopia point (*ED*, PDU), and a weights set (*WM*). Since the results of a given strategy can largely depend on such information, a fair strategy should require as less additional information as possible. The weights should be set by a decision-maker with deep knowledge of the hierarchy among the objectives. In contrast, the reference and the (generalized) utopia points are ordinarily intrinsic to the problem. Despite some common practices (e.g., nadir point) [26], it has been shown that determining a reference point $\mathbf{r}$ for $\mathcal{HV}$ is generally more challenging [18, 26], and a badly defined reference point can lead to inconsistent evaluation results [24]. Indeed, having a reference point slightly different from the nadir point could lead to incongruous evaluation, as experimentally demonstrated by Ishibuchi et al. [19]. Therefore, the utopia point is the most effortlessly additional information that can be exploited for this task. Third, **the strategy should not require to scale the range of the objectives**. Scaling may be needed for strategies whose calculation involves objective blending, i.e. *U-KP*, *ED*, *WM*, and PDU. When the objectives have different scales, the bigger the range of an objective, the bigger its contribution to the selection of a solution. However, the choice of scaling the objectives is left to the system designer. Fourth, **the strategy should be deterministic**. The *U-KP* strategy requires randomly extracting a set of weights from a uniform distribution. This could potentially affect the consistency and reproducibility of results. Fifth, **the strategy should equally promote the solutions despite their position on the Pareto frontier**. The strategies blending the objectives are not biased to select solutions based on particular Pareto frontier regions. This is not true for the $\mathcal{HV}$ strategy that tends to promote the solutions on the concave region of a Pareto frontier.

Final Observations and Calibration. To summarize, none of the strategies own all the properties. However, some considerations can be made. *A-KP* and *U-KP* are characterized by huge drawbacks. The former can be utilized only in contexts considering two objectives. The latter is nondeterministic. Furthermore, none of the techniques is able to select a solution irrespective of its position on the Pareto frontier and to be independent of scaling the objective ranges before calculation simultaneously. In this regard, a system designer could prefer to adopt a technique able to fairly choose a solution despite its position on the Pareto frontier (as done by *U-KP*, *ED*, *WM*, and PDU). Indeed, scaling the objectives can be easily performed with a simple operation such as min/max normalization. Furthermore, this operation is subject to the system designer, who can consider the objectives range in specific applications. Concerning the additional knowledge problem, only *A-KP* and *U-KP* do not need supplementary information. However, as stated before, they are characterized by main drawbacks. Then, such additional knowledge is required. Among the remainder techniques, PDU and *ED* exploit easier-to-define additional material, i.e., the utopia point.

Table 1: Overview of the properties of PDU and other selection strategies. The symbols ✓ (✗,—) indicate that the method has (does not have, could not have) the specified property.

Method	*A-KP*	*U-KP*	$\mathcal{HV}$	*ED*	*WM*	PDU
Suitable With >2 Objectives	✗	✓	✓	✓	✓	✓
No Additional Knowledge	✓	✓	$\mathbf{r}$	f°	$\mathbf{w}$	f°
No Scaling before Calculation	✓	—	✓	—	—	—
Deterministic	✓	✗	✓	✓	✓	✓
Equal Treatment of PF Regions	✗	✓	✗	✓	✓	✓
Calibration	✗	✗	✗	✗	✗	✓

By looking beyond, the proposed PDU allows us to define a utopia point for each sample in the dataset. While the other approaches exploit only aggregated models' performance, PDU opens to a novel "calibrated" way to select one—best Pareto-optimal solution tailored to individual sample characteristics. To the best of our knowledge, this is the first attempt to introduce this kind of feature in the task of Pareto-optimal solutions selection strategy.

From now on, when no confusion arises, we will use *utopia point* to refer also to a *generalized utopia point*.

5 EXPERIMENTAL EVALUATION

We now present an experimental evaluation based on public data that aims at answering the following research questions:

- **RQ1**: How do PDU and other state-of-the-art selection strategies behave w.r.t. the discussed properties? (see Section 4.2)
- **RQ2**: How does the distribution of the points composing the points on the Pareto frontier influence the selection of a solution?
- **RQ3**: How does the calibration feature impact the selection of a solution?

5.1 Experimental Scenarios

Driven by the observation that, in IR and RS settings, the Pareto frontier is populated by points representing aggregated results, we analyze the selection strategies in these two settings.

Information Retrieval Scenario. Concerning the IR scenario, we focus on an ad-hoc search task by dealing with the efficiency / effectiveness / energy-consumption trade-off of query processing in IR systems based on machine-learned ranking models [7]. IR systems heavily exploit supervised techniques for learning document ranking models that are both effective and efficient, i.e., able to retrieve within a limited time budget high-quality documents relevant to users' queries. State-of-the-art learning-to-rank models include ensembles of regression trees trained with gradient boosting algorithms, e.g., LambdaMART [7, 52], and deep neural networks, e.g.,

NeuralNDCG [37]. Since ranking is a complex task and the training datasets are large, the learned models are complex and computationally expensive at inference time. The tight constraints on query response time thus require suitable solutions to provide an optimal trade-off between efficiency and ranking quality [8, 16, 30].

In this scenario, we use the LambdaMART [7, 52] implementation available in LightGBM [23] to train ranking models based on ensembles of regression trees and Neural Networks (NN) trained in Pytorch [36] following the optimization methodology proposed in [33]. The models are trained on `MSN30K` [38], a public and widely-used dataset for learning to rank. The evaluation employs 11 LambdaMART and 5 Neural Networks ranking models, each tested on the 6,306 queries of the `MSN30K` test set. We measure the ranking quality of each model in terms of average nDCG@10 (f_1), and average ranking time (seconds per document) (f_2). For the LambdaMART configurations, we also measure the average energy consumption (Joules per document) (f_3). The average ranking time of each model has been measured by using QuickScorer [30], while energy consumption has been measured by using the Mammut library [13]. Efficiency experiments are performed on a dedicated Intel Xeon CPU E5-2630 v3 clocked at 2.4 GHz in single-thread execution. QuickScorer is compiled using GCC 9.2.1 with the `-O3` option.

In this IR experimental scenario, we focus on selecting the best efficiency/effectiveness trade-off for query processing.

Recommendation Scenario. Concerning the RS scenario, we consider two of the main problems of recommendation algorithms, i.e., the accuracy of the recommendations and the tendency to over-suggest popular items. Often, the ability of RS to provide accurate recommendations is competing with the capability of including long-tail items in such suggestions [32], inducing a trade-off. Hence, we consider two objectives. We compute the Recall@10 (f_1) to measure the accuracy of suggestions and the average percentage of items in the long-tail (APLT) [2] (f_2) to measure to what extent a RS can recommend unpopular items:

$$APLT = \frac{1}{|\mathcal{U}_t|} \sum_{u \in \mathcal{U}_t} \frac{|\{i, i \in (\mathcal{L}_u \cap \Phi)\}|}{|\mathcal{L}_u|}, \tag{13}$$

where $|\mathcal{U}_t|$ is the number of users in the test set, $\mathcal{L}_u$ is the list of recommended items to user u, and Φ is the set of long-tail items. The higher the metric, the higher the number of niche items suggested.

Specifically, we interpret APLT from two perspectives, identifying two experimental scenarios. On the one hand, we assess APLT from provider-side fairness. The provider side fairness can be quantified as the models' ability to expose items to users evenly [1, 2, 51]. Indeed, the over-recommendation of popular items, i.e., the so-called unfairness of popularity bias, may be felt as unfair by providers who get long-tail items under-represented in the suggestions. Hence, in this scenario, the goal is to choose a model that promotes relevant items without affecting niche items' visibility.

In this first RS experimental scenario, we focus on selecting the best recommendation model dealing with multiple objectives.

On the other hand, we evaluate APLT from the final user point of view. Indeed, certain users may prefer to consume popular items, while others niche items. Consequently, exclusively recommending mainstream items would hurt the experience of long-tail users, and vice versa. The approach of calibrated recommendation has shown a valuable solution toward this direction of research [35, 42]. A recommendation list is calibrated concerning popularity when the set of items it covers matches the user's profile in terms of item popularity [3]. Inspired by the concept of popularity-based calibrated recommendation, for each user, we compute the values of the APLT target (f_2) stemming from their popularity profile. To this end, we compute the user-level APLT utopia values using the *weighted combination of mean and standard deviation* method described by Jugovac et al. [20]. We consider the set of users $\mathcal{U}$, the set of items $\mathcal{I}$, and the mean number of transactions T in the training set. For each item $i \in \mathcal{I}$, we assess its popularity pop_i by counting the number of transactions the item is involved in. For each user $u \in \mathcal{U}$, we define the set $\Gamma_u = \{pop_i \mid u \text{ interacted with } i\}$. We quantify the user u popularity tendencies as $pop_u = \alpha \cdot \mu(\Gamma_u) + \beta \cdot \sigma(\Gamma_u)$, where α and β are set to a fixed value of 1 as done in [20], $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation operators, respectively. The higher is pop_u, the most user u has consumed mainstream items in her past interactions. Finally, we normalize pop_u and compute the APLT utopia value for each user:

$$f_2^\circ = h_2(u) = \frac{pop_\Psi - pop_u}{pop_\Psi - pop_\Phi}, \tag{14}$$

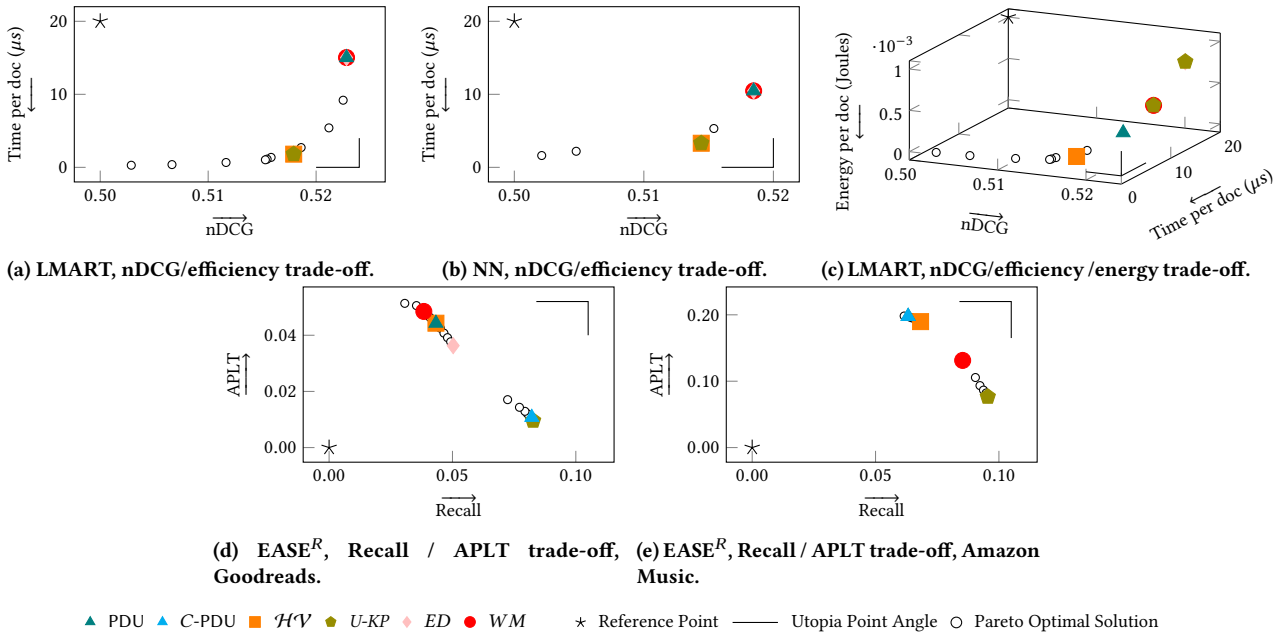
where Φ and Ψ are the sets composed by pop_i values such that i is one of the less and most T consumed items, respectively. With this normalization, the higher is f_2°, the less popular is the user profile.

In this second RS experimental scenario, we show how important a calibrated technique is for choosing the best recommendation model dealing with multiple objectives.

In the two experimental scenarios presented for RS, we exploit the EASE^R recommendation model [43], which works like a shallow autoencoder. This model is characterized by a single hyperparameter to tune, i.e., the L2-norm regularization (λ). Nevertheless, it has been shown that it often outperforms other state-of-the-art recommender systems [4]. Specifically, we explore the hyperparameter λ by training 48 configurations on the book-domain dataset *Goodreads* [48] (18,892 users, 25,475 items, and 1,378,033 transactions) and on the music-domain dataset *Amazon Music* [4] (14,354 users, 10,027 items, and 145,523 transactions). We split the datasets following the 70-10-20 hold-out strategy. Thus, the evaluation of this scenario employs 48 solutions on the objective function space, each tested on the remaining users of the test set (18,070 of *Goodreads*, and 14,354 of *Amazon Music*).

5.2 Experimental Methodology

The different hyperparameter configurations introduced before, for the two IR and RS settings, generate solutions in the objectives function space for each specific experimental scenario. Once the Pareto-optimal solutions that compose the Pareto frontier are identified, we select one by applying PDU and the other selection strategies we analyzed in this work. The selected solutions are then analyzed according to the features introduced in Section 4.2. Moreover, we investigate in detail how the formulation of PDU and its calibration feature influence the choice of the one—best solution by looking at the distribution of points composing that solution. We refer to the reference point and the utopia point with $\mathbf{r}$ and $\boldsymbol{f}^\circ$, respectively. Furthermore, we use the Euclidean distance as $e(\cdot)$ in the formulation of PDU, to have an immediate comparison with *ED*



(a) LMART, nDCG/efficiency trade-off. **(b) NN, nDCG/efficiency trade-off.** **(c) LMART, nDCG/efficiency /energy trade-off.**

(d) EASE^R, Recall / APLT trade-off, Goodreads. **(e) EASE^R, Recall / APLT trade-off, Amazon Music.**

Figure 1: Pareto-optimal solutions for the IR and RS scenarios. The colored shapes represent the best—Pareto-optimal—point selected by the strategies under evaluation.

Table 2: LambdaMART selected solutions for the IR scenario. The objectives are accuracy (*nDCG*), efficiency (*Seconds*), and energy consumption (*Joules*).

Models		Objectives			Selection Strategies									
					Accuracy / Efficiency					Accuracy / Efficiency / Energy				
Trees	Leaves	nDCG ↑	Seconds ↓	Joules ↓	PDU ↓	$\mathcal{HV}$ ↑	*U-KP*↑	*ED* ↓	*WM* ↑	PDU ↓	$\mathcal{HV}$ ↑	*U* ↑	*ED* ↓	*WM* ↑
300	32	0.5179	18.0544×10^{-5}	10.8515×10^{-5}	7.4953	$\mathbf{3.2612 \times 10^{-7}}$	**107**	0.4821	0.1295	7.4960	$\mathbf{2.9236 \times 10^{-10}}$	85	0.4821	0.0863
300	64	0.5212	54.0393×10^{-5}	31.7795×10^{-5}	7.4837	3.0924×10^{-7}	102	0.4788	0.1303	**7.4904**	2.1097×10^{-10}	93	0.4788	0.0868
500	64	0.5225	91.9204×10^{-5}	54.5946×10^{-5}	7.4799	2.4323×10^{-7}	103	0.4775	0.1306	7.4996	1.1044×10^{-10}	**102**	0.4775	**0.0870**
878	64	**0.5228**	150.355×10^{-5}	89.4260×10^{-5}	**7.4768**	1.1328×10^{-7}	98	**0.4772**	**0.1307**	7.5289	0.1198×10^{-10}	**102**	**0.4772**	0.0870

to assess the impact of the points distribution composing a solution. Tables 2, 3, and 4 report the results for the solutions chosen by at least one strategy. For the sake of completeness, the reader may find the complete sets of results in the GitHub repository. The best values for each metric are in bold, while the arrows indicate if better stands for lower ↓ or higher ↑ values.

Experimental settings for the IR scenario. A nadir point cannot be established for the IR scenario because two of the objectives, i.e., efficiency and energy consumption, are not bounded in the opposite direction of the optimization target. For this reason, we define the reference point by slightly worsening the worst values reached by the optimal solutions available. By doing so, we end up setting $\mathbf{r} = (0.5, 0.00002, 0.001)$ for $\mathcal{HV}$. Moreover, we set $\boldsymbol{f}^\circ = (1, 0, 0)$ for *ED*, and for each sample in the dataset in PDU. For what regards *WM*, we equally treat the objectives by setting each weight to 0.5. Finally, in this scenario, we do not apply any normalization to the objective values achieved with the different models.

Experimental settings for the RS scenario. Differently from the IR scenario, a nadir point can be established here because the two objectives under consideration, i.e., Recall and APLT, are bounded. We thus set $\mathbf{r} = (0, 0)$ for $\mathcal{HV}$, and $\boldsymbol{f}^\circ = (1, 1)$ for *ED*. As before, we give equal importance to the objectives in *WM* by setting each weight to 0.5. Concerning PDU, we set $f_1^\circ = 1$ for each sample utopia point as we want all users to have accurate recommendations. Instead, we set $f_2^\circ = 1$ in the first RS experimental scenario, while we compute specific values of f_2° for each user as in Equation (14) in the second RS experimental scenario. Finally, in both RS scenarios, we apply a min-max normalization to the objectives.

We first divide the results discussion according to both IR and RS scenarios for RQ1 and RQ2. Then, we answer RQ3 by exploiting the second RS scenario.

5.3 Performance Comparison (RQ1)

IR scenario. We answer RQ1 by first focusing on the IR scenario. The results for this scenario are summarized in Tables 2 (LambdaMART) and 3 (Neural Networks). The plots in Figures 1a and 1c show the Pareto-optimal points selected by the different techniques for the cases considering two and three objectives regarding the LambdaMART models, respectively. Figure 1b shows the points selected in the case of the Neural Networks models.

Regarding the two-objective case, we observe that the methods blending the objectives (PDU, *ED*, *WM*) select the same Pareto-optimal solution lying on the boundary of the Pareto frontier for both families of models, thus maximizing the accuracy at the cost of efficiency. In contrast, $\mathcal{HV}$ chooses an inner solution of the Pareto frontier in both cases, i.e., more efficient models, that however show a significantly lower performance in terms of nDCG compared to the selection provided by PDU (0.5225 vs. 0.5179 for LambdaMART, and 0.5185 vs. 0.5144 for the Neural Network). It is worth noting that, in this case, no transformation has been applied to the scale of the objectives, and the values of the Pareto solutions for what regards the efficiency scale lead the points to be closer to the utopia value $f_2^\circ = 0$. If a min/max normalization is applied to the objective, PDU still selects the same solution. Another essential implication arising from this analysis is that, in this scenario, we cannot establish the nadir point, making challenging the definition of the reference point. Consequently, this potentially leads to different results based on how we define the reference point. Indeed, as we push the reference point away from the Pareto frontier, $\mathcal{HV}$ selects a boundary solution, as done by PDU. In light of the above results, we observe that if the information related to the nadir point is unavailable, the definition of the reference point can strongly affect the selection of the final solution. Moreover, if the reference point is estimated by looking at the collection of the considered solutions, i.e., by slightly increasing the worst values reached by them, $\mathcal{HV}$ promotes the solution in the middle. Indeed, the definition of the reference point in such a way makes the volume of those solutions, computed as in Equation (2), higher than any other. Thus, $\mathcal{HV}$ unequally considers the remaining points lying on the boundaries of the Pareto frontier. Finally, it is worth highlighting that *U-KP*, although reported in Figures 1a and 1b, is not deterministic. Indeed, by executing this method several times, it may choose different points as the weights of the utility function (see Section 3.1.1) are randomly extracted from a uniform distribution.

Moving to the three-objective formulation of the IR scenario for the LambdaMART models, Figure 1c shows that when introducing the energy consumption objective, the methods tend to choose a more efficient model than the one selected in the two-objectives scenario. As before, PDU and *ED* tend to maximize the accuracy with respect to $\mathcal{HV}$ that still select solutions in the middle. The three-dimensional scenario confirms two behaviors observed in the two-dimensional one. First, the solution selected by $\mathcal{HV}$ depends on the chosen reference point since it is not possible to define a nadir point. Second, *U-KP* still exhibits a non-deterministic behavior.

Finally, we claim that PDU and *ED* perform the most convenient selection from a qualitative perspective. By looking at Tables 2 and 3, we see that they choose the models with higher values of nDCG for all IR cases. Indeed, both efficiency and energy consumption objectives are closer to their respective utopia values. This means that more complex models, chosen by PDU and *ED*, guarantee considerable improvement in ranking accuracy at a small reduction of efficiency and energy consumption. Conversely, $\mathcal{HV}$ chooses models that exhibit a considerable decrease in terms of nDCG.

RS scenario. For the first RS experimental scenario, we report the results achieved in Table 4 for the Goodreads dataset (Figure 1d) and for the Amazon Music dataset (Figure 1e). For both datasets, we

Table 3: Neural Networks selected solutions in the IR scenario. The objectives are accuracy (*nDCG*) and efficiency (*Seconds*).

Models				Objectives			Selection Strategies			
L1	L2	L3	L4	nDCG ↑	Seconds ↓	PDU ↓	$\mathcal{HV}$ ↑	*U-KP*↑	*ED* ↓	*WM* ↑
100	50	50	10	0.5144	3.3003×10^{-6}	7.5069	$\mathbf{2.4099 \times 10^{-7}}$	**221**	0.4856	0.1286
200	100	100	50	**0.5185**	1.0476×10^{-5}	**7.4959**	1.7598×10^{-7}	204	**0.4815**	**0.1296**

Table 4: EASE^R selected solutions (for Goodreads and Amazon Music) in the RS scenario with *Recall* and *APLT* objectives. For *APLT*, the higher the better refers to the provider side.

Models	Objectives				Selection Strategies			
λ	Recall ↑	APLT ↑*	PDU ↓	C-PDU ↓	$\mathcal{HV}$ ↑	*U-KP*↑	*ED* ↓	*WM* ↑
Goodreads								
0.3	0.0384	0.0485	10.4113	10.0898	0.1861×10^{-2}	55	0.8546	**0.2699**
0.5	0.0433	0.0443	**10.4066**	10.0829	$\mathbf{0.1919 \times 10^{-2}}$	16	0.7761	0.2686
1	0.0503	0.0363	10.4098	10.0819	0.1826×10^{-2}	0	**0.7191**	0.2546
60	0.0822	0.0108	10.4126	**10.0706**	0.0885×10^{-2}	86	0.9651	0.2556
90	0.0827	0.0096	10.4134	10.0711	0.0791×10^{-2}	**101**	0.9938	0.2510
Amazon Music								
0.3	0.0632	0.1976	**10.0104**	**9.8604**	0.1249×10^{-1}	79	0.9524	0.2608
1	0.0683	0.1898	10.0147	9.8628	$\mathbf{0.1295 \times 10^{-1}}$	49	0.8074	0.2819
10	0.0853	0.1313	10.0784	9.9160	0.1120×10^{-1}	4	**0.6177**	**0.2896**
80	0.0955	0.0766	10.1268	9.9570	0.0731×10^{-1}	**89**	0.9780	0.2542

notice that two well-separated clusters characterize the Pareto frontier. On the one hand, in Goodreads the EASE^R configurations with lower L2 norm (λ) values, which belong to the top-center cluster, account for the accommodation of the objectives. In contrast, the second cluster (bottom-right), i.e., λ between 10 and 100 in Table 4, maximizes Recall at the expense of the exposure of the items (lower values of APLT). On the other hand, in Amazon Music, these two clusters of configurations follow the opposite behavior. On the one side, the configurations with λ between 0.2 and 1 maximize APLT at the detriment of Recall (top-left cluster). On the other side, the remaining configurations do not promote either Recall or APLT (bottom-right cluster). In this scenario, $\mathcal{HV}$ suffers less from the problem of promoting solutions in the center of the Pareto frontier. Indeed, differently from the IR scenario, here it is possible to define the nadir point as a reference point because we know the lowest bounds (0 for both APLT and Recall). Consequently, even though $\mathcal{HV}$ selects an inner solution in the Goodreads case, it chooses a point that tends to maximize APLT for the Amazon Music dataset. PDU follows the behaviour of $\mathcal{HV}$ when selecting the solutions for both datasets. By considering that it selects an outer point of the Pareto frontier in the IR scenario, this endorses the ability of PDU to equally promote the available solutions despite their positioning on the Pareto frontier. *WM* and *ED* select a solution belonging to the top-center cluster in Goodreads and to the bottom-right cluster in Amazon Music, thus enhancing the trade-off between accuracy measured in terms of Recall and items exposure in both cases. Finally, *U-KP* still exhibits a nondeterministic performance.

To answer RQ1 we conclude observing that PDU overcomes some limitations of $\mathcal{HV}$ and U-KP competitors. Indeed, PDU selects one—best—Pareto-optimal solution regardless of its position on the Pareto

frontier in a deterministic way. Moreover, it exploits the concept of Utopia point as additional information. Such a concept is more convenient to use than the reference point used in $\mathcal{HV}$, since, depending on the problem addressed, the nadir point is difficult to be defined.

5.4 Impact of the Points Distribution (RQ2)

We now answer RQ2 by investigating the impact on selecting the distribution of the points that compose a solution on the Pareto frontier. Indeed, PDU is the only strategy considering these points in a more fine-grained resolution. This analysis is done on both the IR (Tables 2 and 3) and RS (Table 4) scenarios. To this end, we remember that we have set $e(\cdot)$ as the Euclidean Distance in the formulation of PDU (Equation (3)). Hence, even if both PDU and *ED* rely on the Euclidean distance, they work differently in the two experimental scenarios. This observation provides insights on the impact of the points distribution on the selection.

RS scenario. PDU and *ED* choose different solutions for both RS datasets. In this regard, the user data points' distribution in the objective function space plays a crucial role. To confirm this fact, we compute the users points' mean Euclidean distances to the utopia point of both solutions. Results confirm that the EASE^R configuration selected by PDU has a lower value of average Euclidean distance, i.e., 1.3498 for $\lambda = 0.5$, w.r.t. the configuration chosen by *ED*, i.e., 1.352 for $\lambda = 1$. The same impact is observed regarding the Amazon Music dataset. Here, PDU and *ED* select different configuration models having $\lambda = 0.3$ and $\lambda = 10$, respectively. As before, the EASE^R configuration selected by PDU ($\lambda = 0.3$) has a lower value of average Euclidean distance, i.e., 1.2361 than the configuration chosen by *ED* ($\lambda = 10$), i.e., 1.279.

IR scenario. Concerning the IR two-objectives cases, PDU and *ED* choose the same solution for both LambdaMART and Neural Networks models. When introducing energy consumption as the third objective for the LambdaMART models, *ED* still selects the same configuration with 878 trees and 64 leaves. Conversely, PDU chooses a more efficient model (300 trees and 64 leaves). Once more, the query points' mean Euclidean distances to the common utopia point of the model selected by PDU are lower than the ones of the model chosen by *ED* (0.4813 vs. 0.4945).

To conclude, the answer to RQ2 is that the distribution of the points composing a solution with respect to a common utopia point has a significant impact on the final selection. This is an important fact, as it paves the way to defining selection strategies that take the distribution of the points into account while performing a selection that can be done in a more—fine-grained—sample-level way.

5.5 Impact of Calibration on the Selection (RQ3)

Finally, we analyze the impact of the calibration introduced for PDU using the second RS scenario, where we aim to tailor the selection according to the users' item popularity tastes. To this end, we assess the selection performed by Calibrated-PDU (C-PDU).

Starting from the Amazon Music dataset, the average of the APLT utopia values computed with Equation (14) (0.83) reveals that the dataset's users generally prefer less popular items. Indeed, C-PDU selects the EASE^R model with $\lambda = 0.3$. This solution lies on the top-left cluster of Figure 1e, by maximizing APLT with a loss of Recall. In this case, C-PDU behaves similarly to PDU and $\mathcal{HV}$. Moving to the Goodreads dataset, it is characterized by users with more mainstream tastes, since the average of the APLT utopia values is equal to 0.65. Surprisingly, C-PDU is the only strategy among the ones tested selecting a model configuration belonging to the bottom-right cluster in Figure 1d where the solutions achieve higher accuracy values without promoting APLT and following the mainstream users tastes — along with *U-KP* that, however, has a non-deterministic behavior. These experimental results already qualitatively show the impact of defining a utopia point for each user on the final selection, since C-PDU is the only strategy to capture the users' popularity profiles for both datasets. We deepen the analysis further by considering the model configurations chosen by PDU and C-PDU for Goodreads, i.e., $\lambda = 0.5$ and $\lambda = 60$, respectively. We observe that, although the model with $\lambda = 0.5$ performs better on average APLT, the model with $\lambda = 60$ has a lower variance of the mean absolute error (0.036 for $\lambda = 60$ vs. 0.039 for $\lambda = 0.5$) between the utopia values and the model performance values for each user. This indicates that C-PDU selects the model that generally follows better the users' popularity profile. In addition, this model provides more accurate recommendations on average. Hence, C-PDU chooses the model that performs better in terms of accuracy and also tailors the popular tastes of the users.

To conclude, the answer to RQ3 is that the calibration feature of PDU allows dealing with the ideal targets for each sample. This confirms that calibration is a viable way to move the selection of the Pareto-optimal solution to a more fine-grained resolution that can lead to significantly different choices in terms of the trade-off selected.

6 CONCLUSIONS AND FUTURE WORK

In this work, we proposed PDU, a novel, theoretically-justified *post-hoc* technique to select one—best—Pareto-optimal solution among the ones lying in the Pareto frontier in search and recommendation scenarios. To our knowledge, PDU is the only selection technique in the literature that can be "calibrated", i.e., it can choose the best Pareto-optimal solution based on ideal targets expressed on single queries or users. We comprehensively compared the properties of PDU with those of competitor techniques. We conducted an extensive experimental evaluation focusing on both IR and RS scenarios, showing that the formulation and the calibration feature of PDU have a notable impact on the solution's selection. In the future, we will explore PDU by exploiting other distance metrics (e.g., Chebyshev and Manhattan). Moreover, it could be interesting to perform online A/B tests to assess the impact of the calibrated selection. Finally, this work could open to the formulation of a new loss function based on the PDU derivation, to directly train a ranking model on multiple objectives simultaneously.

Acknowledgements. This research has been partially funded by: Secure Safe Apulia, Casa delle Tecnologie Emergenti Comune di Matera, LUTECH DIGITALE 4.0, OVS Fashion Retail Reloaded, CT_FINCONS_III, KOINÈ, PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme and by the Horizon Europe RIA "Extreme Food Risk Analytics" (EFRA), grant agreement n. 101093026.

REFERENCES

[1] Himan Abdollahpouri, Gediminas Adomavicius, Robin Burke, Ido Guy, Dietmar Jannach, Toshihiro Kamishima, Jan Krasnodebski, and Luiz Augusto Pizzato. 2020. Multistakeholder recommendation: Survey and research directions. *User Model. User Adapt. Interact.* 30, 1 (2020), 127–158. https://doi.org/10.1007/s11257-019-09256-1

[2] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems, RecSys 2017, Como, Italy, August 27-31, 2017*, Paolo Cremonesi, Francesco Ricci, Shlomo Berkovsky, and Alexander Tuzhilin (Eds.). ACM, 42–46. https://doi.org/10.1145/3109859.3109912

[3] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2019. Managing Popularity Bias in Recommender Systems with Personalized Re-Ranking. In *Proceedings of the Thirty-Second International Florida Artificial Intelligence Research Society Conference, Sarasota, Florida, USA, May 19-22 2019*, Roman Barták and Keith W. Brawner (Eds.). AAAI Press, 413–418. https://aaai.org/ocs/index.php/FLAIRS/FLAIRS19/paper/view/18199

[4] Vito Walter Anelli, Alejandro Bellogín, Tommaso Di Noia, Dietmar Jannach, and Claudio Pomo. 2022. Top-N Recommendation Algorithms: A Quest for the State-of-the-Art. In *UMAP '22: 30th ACM Conference on User Modeling, Adaptation and Personalization, Barcelona, Spain, July 4 - 7, 2022*, Alejandro Bellogín, Ludovico Boratto, Olga C. Santos, Liliana Ardissono, and Bart P. Knijnenburg (Eds.). ACM, 121–131. https://doi.org/10.1145/3503252.3531292

[5] Jürgen Branke, Kalyanmoy Deb, Henning Dierolf, and Matthias Osswald. 2004. Finding Knees in Multi-objective Optimization. In *Parallel Problem Solving from Nature - PPSN VIII, 8th International Conference, Birmingham, UK, September 18-22, 2004, Proceedings (Lecture Notes in Computer Science, Vol. 3242)*, Xin Yao, Edmund K. Burke, José Antonio Lozano, Jim Smith, Juan Julián Merelo Guervós, John A. Bullinaria, Jonathan E. Rowe, Peter Tiňo, Ata Kabán, and Hans-Paul Schwefel (Eds.). Springer, 722–731. https://doi.org/10.1007/978-3-540-30217-9_73

[6] Jürgen Branke, Kalyanmoy Deb, Kaisa Miettinen, and Roman Slowinski (Eds.). 2008. *Multiobjective Optimization, Interactive and Evolutionary Approaches [outcome of Dagstuhl seminars]*. Lecture Notes in Computer Science, Vol. 5252. Springer. https://doi.org/10.1007/978-3-540-88908-3

[7] Christopher JC Burges. 2010. From ranknet to lambdarank to lambdamart: An overview. *Learning* 11, 23-581 (2010), 81.

[8] Gabriele Capannini, Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, and Nicola Tonellotto. 2016. Quality Versus Efficiency in Document Scoring with Learning-to-rank Models. *Information Processing Management* 52, 6 (Nov. 2016), 1161–1177.

[9] David Carmel, Elad Haramaty, Arnon Lazerson, and Liane Lewin-Eytan. 2020. Multi-Objective Ranking Optimization for Product Search Using Stochastic Label Aggregation. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen (Eds.). ACM / IW3C2, 373–383. https://doi.org/10.1145/3366423.3380122

[10] Na Dai, Milad Shokouhi, and Brian D Davison. 2011. Learning to rank for freshness and relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 95–104.

[11] Onkar Dalal, Srinivasan H. Sengemedu, and Subhajit Sanyal. 2012. Multi-Objective Ranking of Comments on Web. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12)*. Association for Computing Machinery, New York, NY, USA, 419–428. https://doi.org/10.1145/2187836.2187894

[12] Kalyanmoy Deb and Shivam Gupta. 2011. Understanding knee points in bicriteria problems and their implications as preferred solution principles. *Engineering Optimization* 43, 11 (2011), 1175–1204. https://doi.org/10.1080/0305215X.2010.548863 arXiv:https://doi.org/10.1080/0305215X.2010.548863

[13] Daniele De Sensi, Massimo Torquati, and Marco Danelutto. 2017. Mammut: High-level management of system knobs and sensors. *SoftwareX* 6 (2017), 150–154. https://doi.org/10.1016/j.softx.2017.06.005

[14] M. Fleischer. 2003. The Measure of Pareto Optima. In *Evolutionary Multi-Criterion Optimization, Second International Conference, EMO 2003, Faro, Portugal, April 8-11, 2003, Proceedings (Lecture Notes in Computer Science, Vol. 2632)*, Carlos M. Fonseca, Peter J. Fleming, Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele (Eds.). Springer, 519–533. https://doi.org/10.1007/3-540-36970-8_37

[15] Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. 2022. Toward Pareto Efficient Fairness-Utility Trade-off in Recommendation through Reinforcement Learning. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 316–324. https://doi.org/10.1145/3488560.3498487

[16] Veronica Gil-Costa, Fernando Loor, Romina Molina, Franco Maria Nardini, Raffaele Perego, and Salvatore Trani. 2022. Ensemble Model Compression for Fast and Energy-Efficient Ranking on FPGAs. In *Advances in Information Retrieval - 44th European Conference on IR Research, ECIR 2022, Stavanger, Norway, April 10-14, 2022, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 13185)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer, 260–273. https://doi.org/10.1007/978-3-030-99736-6_18

[17] Haimes, Lasdon, and Wismer. 1971. On a Bicriterion Formulation of the Problems of Integrated System Identification and System Optimization. *IEEE Transactions on Systems, Man, and Cybernetics* SMC-1, 3 (1971), 296–297. https://doi.org/10.1109/TSMC.1971.4308298

[18] Michael Pilegaard Hansen and Andrzej Jaszkiewicz. 1994. *Evaluating the quality of approximations to the non-dominated set.* IMM, Department of Mathematical Modelling, Technical Universityof Denmark.

[19] Hisao Ishibuchi, Ryo Imada, Yu Setoguchi, and Yusuke Nojima. 2018. How to Specify a Reference Point in Hypervolume Calculation for Fair Performance Comparison. *Evol. Comput.* 26, 3 (2018). https://doi.org/10.1162/evco_a_00226

[20] Michael Jugovac, Dietmar Jannach, and Lukas Lerche. 2017. Efficient optimization of multiple recommendation quality factors according to individual user tendencies. *Expert Syst. Appl.* 81 (2017), 321–331. https://doi.org/10.1016/j.eswa.2017.03.055

[21] Changsung Kang, Xuanhui Wang, Yi Chang, and Belle Tseng. 2012. Learning to Rank with Multi-Aspect Relevance for Vertical Search. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. Association for Computing Machinery, New York, NY, USA, 453–462. https://doi.org/10.1145/2124295.2124350

[22] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2017. On Application of Learning to Rank for E-Commerce Search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. Association for Computing Machinery, New York, NY, USA, 475–484. https://doi.org/10.1145/3077136.3080838

[23] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 3149–3157.

[24] Joshua D. Knowles and David Corne. 2003. Properties of an adaptive archiving algorithm for storing nondominated vectors. *IEEE Trans. Evol. Comput.* 7, 2 (2003), 100–116. https://doi.org/10.1109/TEVC.2003.810755

[25] Charles R. Leake. 2001. Multicriterion Decision in Management: Principles and Practice. *J. Oper. Res. Soc.* 52, 5 (2001), 603. https://doi.org/10.1057/palgrave.jors.2601200

[26] Miqing Li and Xin Yao. 2019. Quality Evaluation of Solution Sets in Multiobjective Optimisation: A Survey. *ACM Comput. Surv.* 52, 2 (2019), 26:1–26:38. https://doi.org/10.1145/3300148

[27] M. Lightner and S. Director. 1981. Multiple criterion optimization for the design of electronic circuits. *IEEE Transactions on Circuits and Systems* 28, 3 (1981), 169–179. https://doi.org/10.1109/TCS.1981.1084969

[28] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems, RecSys 2019, Copenhagen, Denmark, September 16-20, 2019*, Toine Bogers, Alan Said, Peter Brusilovsky, and Domonkos Tikk (Eds.). ACM, 20–28. https://doi.org/10.1145/3298689.3346998

[29] Bo Long, Jiang Bian, Anlei Dong, and Yi Chang. 2012. Enhancing Product Search by Best-Selling Prediction in e-Commerce. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, Hawaii, USA) *(CIKM '12)*. Association for Computing Machinery, New York, NY, USA, 2479–2482. https://doi.org/10.1145/2396761.2398671

[30] Claudio Lucchese, Franco Maria Nardini, Salvatore Orlando, Raffaele Perego, Nicola Tonellotto, and Rossano Venturini. 2015. QuickScorer: A Fast Algorithm to Rank Documents with Additive Ensembles of Regression Trees. In *Proc. ACM SIGIR*. 73–82.

[31] R. Marler and Jasbir Arora. 2004. Survey of Multi-Objective Optimization Methods for Engineering. *Structural and Multidisciplinary Optimization* 26 (04 2004), 369–395. https://doi.org/10.1007/s00158-003-0368-6

[32] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Yashar Deldjoo. 2022. CP-Fair: Personalized Consumer and Producer Fairness Re-ranking for Recommender Systems. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, Enrique Amigó, Pablo Castells, Julio Gonzalo, Ben Carterette, J. Shane Culpepper, and Gabriella Kazai (Eds.). ACM, 770–779. https://doi.org/10.1145/3477495.3531959

[33] Franco Maria Nardini, Cosimo Rulli, Salvatore Trani, and Rossano Venturini. 2022. Distilled Neural Networks for Efficient Learning to Rank. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[34] Tommaso Di Noia, Jessica Rosati, Paolo Tomeo, and Eugenio Di Sciascio. 2017. Adaptive multi-attribute diversity for recommender systems. *Inf. Sci.* 382-383 (2017), 234–253. https://doi.org/10.1016/j.ins.2016.11.015

[35] Jinoh Oh, Sun Park, Hwanjo Yu, Min Song, and Seung-Taek Park. 2011. Novel Recommendation Based on Personal Popularity Tendency. In *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, Diane J. Cook, Jian Pei, Wei Wang, Osmar R. Zaïane, and Xindong Wu

(Eds.). IEEE Computer Society, 507–516. https://doi.org/10.1109/ICDM.2011.110

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 8024–8035. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[37] Przemyslaw Pobrotyn and Radoslaw Bialobrzeski. 2021. NeuralNDCG: Direct Optimisation of a Ranking Metric via Differentiable Relaxation of Sorting. *CoRR* abs/2102.07831 (2021). arXiv:2102.07831 https://arxiv.org/abs/2102.07831

[38] Tao Qin and Tie-Yan Liu. 2013. Introducing LETOR 4.0 Datasets. *CoRR* abs/1306.2597 (2013). arXiv:1306.2597 http://arxiv.org/abs/1306.2597

[39] Marco Túlio Ribeiro, Anísio Lacerda, Adriano Veloso, and Nivio Ziviani. 2012. Pareto-efficient hybridization for multi-objective recommender systems. In *Sixth ACM Conference on Recommender Systems, RecSys '12, Dublin, Ireland, September 9-13, 2012*, Padraig Cunningham, Neil J. Hurley, Ido Guy, and Sarabjot Singh Anand (Eds.). ACM, 19–26. https://doi.org/10.1145/2365952.2365962

[40] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2015. Search Result Diversification. *Foundations and Trends® in Information Retrieval* 9, 1 (2015), 1–90. https://doi.org/10.1561/1500000040

[41] Dusan Stamenkovic, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. 2022. Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. In *WSDM '22: The Fifteenth ACM International Conference on Web Search and Data Mining, Virtual Event / Tempe, AZ, USA, February 21 - 25, 2022*, K. Selcuk Candan, Huan Liu, Leman Akoglu, Xin Luna Dong, and Jiliang Tang (Eds.). ACM, 957–965. https://doi.org/10.1145/3488560.3498471

[42] Harald Steck. 2018. Calibrated recommendations. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 154–162. https://doi.org/10.1145/3240323.3240372

[43] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 3251–3257. https://doi.org/10.1145/3308558.3313710

[44] Krysta M. Svore, Maksims N. Volkovs, and Christopher J.C. Burges. 2011. Learning to Rank with Multiple Objective Functions. In *Proceedings of the 20th International Conference on World Wide Web* (Hyderabad, India) *(WWW '11)*. Association for Computing Machinery, New York, NY, USA, 367–376. https://doi.org/10.1145/1963405.1963459

[45] Joost van Doorn, Daan Odijk, Diederik M. Roijers, and Maarten de Rijke. 2016. Balancing Relevance Criteria through Multi-Objective Optimization. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 769–772. https://doi.org/10.1145/2911451.2914708

[46] Saul Vargas and Pablo Castells. 2011. Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 2011 ACM Conference on Recommender Systems, RecSys 2011, Chicago, IL, USA, October 23-27, 2011*, Bamshad Mobasher, Robin D. Burke, Dietmar Jannach, and Gediminas Adomavicius (Eds.). ACM, 109–116. https://dl.acm.org/citation.cfm?id=2043955

[47] Christian von Lücken, Benjamín Barán, and Carlos A. Brizuela. 2014. A survey on multi-objective evolutionary algorithms for many-objective problems. *Comput. Optim. Appl.* 58, 3 (2014), 707–756. https://doi.org/10.1007/s10589-014-9644-1

[48] Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-Grained Spoiler Detection from Large-Scale Review Corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 2605–2610. https://doi.org/10.18653/v1/p19-1248

[49] Lidan Wang, Paul N. Bennett, and Kevyn Collins-Thompson. 2012. Robust Ranking Models via Risk-Sensitive Optimization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Portland, Oregon, USA) *(SIGIR '12)*. Association for Computing Machinery, New York, NY, USA, 761–770. https://doi.org/10.1145/2348283.2348385

[50] Shanfeng Wang, Maoguo Gong, Haoliang Li, and Junwei Yang. 2016. Multi-objective optimization for long tail recommendation. *Knowl. Based Syst.* 104 (2016), 145–155. https://doi.org/10.1016/j.knosys.2016.04.018

[51] Haolun Wu, Chen Ma, Bhaskar Mitra, Fernando Diaz, and Xue Liu. 2022. Multi-FR: A Multi-objective Optimization Framework for Multi-stakeholder Fairness-aware Recommendation. In *Transactions on Information Systems (TOIS)*. ACM.

[52] Q. Wu, C.J.C. Burges, K.M. Svore, and J. Gao. 2010. Adapting boosting for information retrieval measures. *Information Retrieval* (2010).

[53] Yong Zheng and David (Xuejun) Wang. 2022. A survey of recommender systems with multi-objective optimization. *Neurocomputing* 474 (2022), 141–153. https://doi.org/10.1016/j.neucom.2021.11.041

[54] Eckart Zitzler, Dimo Brockhoff, and Lothar Thiele. 2007. The Hypervolume Indicator Revisited: On the Design of Pareto-compliant Indicators Via Weighted Integration. In *Evolutionary Multi-Criterion Optimization, 4th International Conference, EMO 2007, Matsushima, Japan, March 5-8, 2007, Proceedings (Lecture Notes in Computer Science, Vol. 4403)*, Shigeru Obayashi, Kalyanmoy Deb, Carlo Poloni, Tomoyuki Hiroyasu, and Tadahiko Murata (Eds.). Springer, 862–876. https://doi.org/10.1007/978-3-540-70928-2_64