A Survey on Rank Aggregation

Siyi Wang[1] , **Qi Deng**[1] , **Shiwei Feng** [1,3,4] , **Hong Zhang**[2,3] and **Chao Liang**[1,2,3,4] *
[1]School of Computer Science, Wuhan University
[2]School of Cyber Science and Engineering, Wuhan University
[3]National Engineering Research Center for Multimedia Software (NERCMS)
[4]Hubei Key Laboratory of Multimedia and Network Communication Engineering
{2021302111095, 2021302111226, waii2022, 2018302180055, cliang}@whu.edu.cn

Abstract

Rank aggregation (RA), the technique of combining multiple basic rankings into a consensus one, plays an important role in social choice, bioinformatics, information retrieval, metasearch, and recommendation systems. Although recent years have witnessed remarkable progress in RA, the absence of a systematic overview motivates us to conduct a comprehensive survey that includes both classic algorithms and the latest advances in RA study. Specifically, we first discuss the challenges of RA research, then present a systematic review with a fine-grained taxonomy to introduce representative algorithms in unsupervised RA, supervised RA, as well as the previously overlooked semi-supervised RA. Within each category, we not only summarize the common ideas of similar methods, but also discuss their strengths and weaknesses. Particularly, to investigate the performance difference between different types of RA methods, we conduct the largest scale of comparative evaluation to date of 27 RA methods on 7 public datasets from person re-identification, recommendation systems, bioinformatics and social choice. Finally, we raise two open questions in the current RA research and make our comments about future trends in the context of the latest research progress.

1 Introduction

Rank aggregation (RA), a fundamental technique to combine multiple basic rankings from various rankers into a single consensus one, has wide applications in society, economy, and science [Akritidis *et al.*, 2022]. RA was originally designed to solve the problem of inconsistent voting for political candidates [Borda, 1781]. Since then, numerous research efforts have been devoted to addressing issues related to world university rankings in social choice [Zhang *et al.*, 2021] [Feng *et al.*, 2023], gene sequences combination in bioinformatics [Li *et al.*, 2019] [Wang *et al.*, 2022], detection results aggregation in spam filtering [Liu *et al.*, 2007], and search results fusion in document retrieval [Farah and Vanderpooten, 2007] [Wu, 2013], person re-identification (re-ID) [Yu *et al.*, 2020] [Huang *et al.*, 2022], metasearch [Renda and Straccia, 2003] [Desarkar *et al.*, 2016] and recommendation systems [Oliveira *et al.*, 2020].

Because of the broad interests and fundamental importance of the RA technique, several inspiring review articles [Lin, 2010] [Li *et al.*, 2019] [Oliveira *et al.*, 2020] [Wang *et al.*, 2022] have been published in this field. As the earliest review, [Lin, 2010] provides a systematic overview of unsupervised RA literature at the time with a special focus on biological applications. To investigate the performance, the author conducts an experimental evaluation of 9 unsupervised RA methods on a cancer gene dataset. Thereafter, [Li *et al.*, 2019] and [Wang *et al.*, 2022] further summarize the latest unsupervised RA methods in bioinformatics, and evaluate methods' performance in a series of larger genomic experiments. Meanwhile, [Oliveira *et al.*, 2020] exclusively focuses on RA studies in the field of recommendation systems, and conducts comparative experiments of RA methods in movie recommendation. In general, existing surveys tend to review RA works in a specific field, and primarily concentrate on unsupervised RA methods. However, recent RA studies have achieved significant progress in novel approaches and applications, which have not been thoroughly reviewed in time. Therefore, we argue that conducting a systematic review of classical algorithms and recent advances is urgently necessary to provide up-to-date guidance for future research.

Compared to previous RA surveys, our work makes the following differences and contributions:

- **A systematic and up-to-date overview.** We provide the first comprehensive review of existing RA techniques, including classic methods in not only unsupervised and supervised RA but also semi-supervised RA which is ignored by previous RA surveys. Moreover, we discuss state-of-the-art progress in research and applications.
- **A fine-grained taxonomy.** Based on the types of studied relationships, we further subdivide unsupervised RA into item-item, item-ranking, and ranking-ranking methods. Based on the levels of data granularities, we further subdivide supervised RA into point-wise, pair-wise, and list-wise methods.
- **The largest evaluations across multiple fields.** We

*Corresponding author.

conduct, for the first time as far as we know, the largest scale of experimental evaluation, including 27 RA methods on 7 public datasets from person re-ID, recommendation systems, bioinformatics, and social choice.

- **Open questions and future perspectives.** We discuss two open questions in the current RA research and application, and share our perspective about future trends, *i.e.*, active learning in interactive RA and RA-based adversarial defense, in the context of the state-of-the-art research progress.
- **Open-source algorithm RA-Lib toolkit with public benchmarks.** We integrate all RA algorithm code experimentally evaluated in this paper together with relevant benchmark datasets into a toolbox, which is released open-source on GitHub[1] for researchers.

2 Preliminary

In this section, we first discuss key challenges in RA study. On this basis, we give a comprehensive and fine-grained taxonomy of RA methods. Then, we define the problem of RA and explain the basic notation used throughout this article.

2.1 Challenges

RA aims to combine basic rankings from multiple rankers. Early RA strategies, like average fusion or heuristic-based fusion [Borda, 1781] [Reilly, 2002] are unsupervised methods; their core challenge is to effectively determine the quality of different basic rankings without external supervision, in order to produce an optimal fusion result. Although unsupervised methods are easy to implement and prevalent in RA research, their performance is difficult to improve due to the lack of supervision [Liu *et al.*, 2007].

To address this drawback, researchers turn to introduce external supervision to guide the RA process [Pujari and Kanawati, 2012] [Wang *et al.*, 2013]. Provided with high-quality training data, supervised RA methods can generally achieve better results than their unsupervised counterparts; however, the challenge posed by the high cost of data collection and labeling also limits the performance of supervised RA methods.

Similar to general machine learning problems, the trade-off between low-cost and high-performance persists in RA research. Unsupervised and supervised methods essentially make an either-or choice, where the former prioritizes low-cost while the latter prioritizes high-performance, each with inherent drawbacks in practice. To overcome this dilemma, a more realistic approach and also a more difficult challenge is to find a well-proportioned balance between cost and performance.

2.2 Taxonomy

Based on the above discussion, we propose a systematic and fine-grained taxonomy that reflects the overall landscape of RA study in Figure 1. We generally divide the existing RA methods into unsupervised, supervised, and semi-supervised methods.

[1]https://github.com/nercms-mmap/RankAggregation-Lib

Unsupervised RA has evolved into the majority of RA methods during the development of RA. Numerous unsupervised RA methods have their special fusion strategies. Based on the types of studied relationships, unsupervised RA methods are subdivided into item-item, item-ranking and ranking-ranking methods. Moreover, item-ranking methods, based on the differences in the underlying mathematical models, can be subdivided into probabilistic model and statistics-based methods.

In supervised RA, we categorize supervised RA methods into three classes: point-wise, pair-wise, and list-wise methods, based on the levels of data granularities. The existing list-wise methods can be further subdivided into metric-driven, evolutionary, neural network-based, and interactive methods, according to their methodological principles or characteristics.

Semi-supervised RA methods are not further subdivided because of the limited amount of related work.

2.3 Problem Formulation

Given a set of M items to be sorted $\mathcal{U} = \{u_1, u_2, ..., u_M\}$, we define a basic ranking $R = \{u_i \geq u_j \geq \ldots \geq u_h\}$ as an ordered list of items, where $i, j, h \in \{1, \ldots, M\}$ and $i \neq j \neq h$. Note that R does not necessarily rank all items in $\mathcal{U}$. We denote by $R^t(u_i)$ the position (or rank) of item u_i under ranking R^t. If $R^t(u_i) < R^t(u_j)$, we say u_i is more relevant than u_j in R^t. For a set of N basic rankings $\mathcal{R} = \{R^1, R^2, ..., R^N\}$, the objective of RA is to find a function f to combine these basic rankings in $\mathcal{R}$ into a consensus ranking R^*, *i.e.*, $R^* = f(\mathcal{R})$. For RA methods that aim to calculate the final score of items, we define S as the scoring function and $S(u_i)$ as the score of item u_i.

3 Unsupervised RA Methods

3.1 Item-Item Methods

The item-item methods investigate the preference relationship between items to produce the final result. Condorcet [de Condorcet, 1785] proposes a Condorcet procedure based on pairwise comparisons of items, where one item beats, ties, or loses to the other item. The score of an item u_i in the Condorcet procedure is determined by the number of its "victories" against all the other items in basic rankings. A victory for u_i is achieved if the majority of the basic rankings rank u_i higher than any other item $u_{j \neq i}$. For example, for a set of items $\{u_1, u_2, u_3\}$, the Condorcet procedure conducts pairwise comparisons among these items: if u_1 beats u_2, u_1 beats u_3 and u_3 beats u_2, then the score is $\{2, 0, 1\}$, so we get the preference among these items and fuse these preference into a ranking $R = \{u_1 \geq u_3 \geq u_2\}$.

[Saari and Merlin, 1996] computes Copeland scores of items based on the disparity between the count of items they surpass and the count of items they are surpassed by:

$$S_{\text{Copeland}}(u_i) = \sum_{R^t \in \mathcal{R}} \left|\left\{u_j \in R^t : R^t(u_i) < R^t(u_j)\right\}\right| - \left|\left\{u_s \in R^t : R^t(u_s) < R^t(u_i)\right\}\right|. \quad (1)$$

where $|\cdot|$ returns the set cardinality. The consensus ranking is generated by arranging the items according to their Copeland

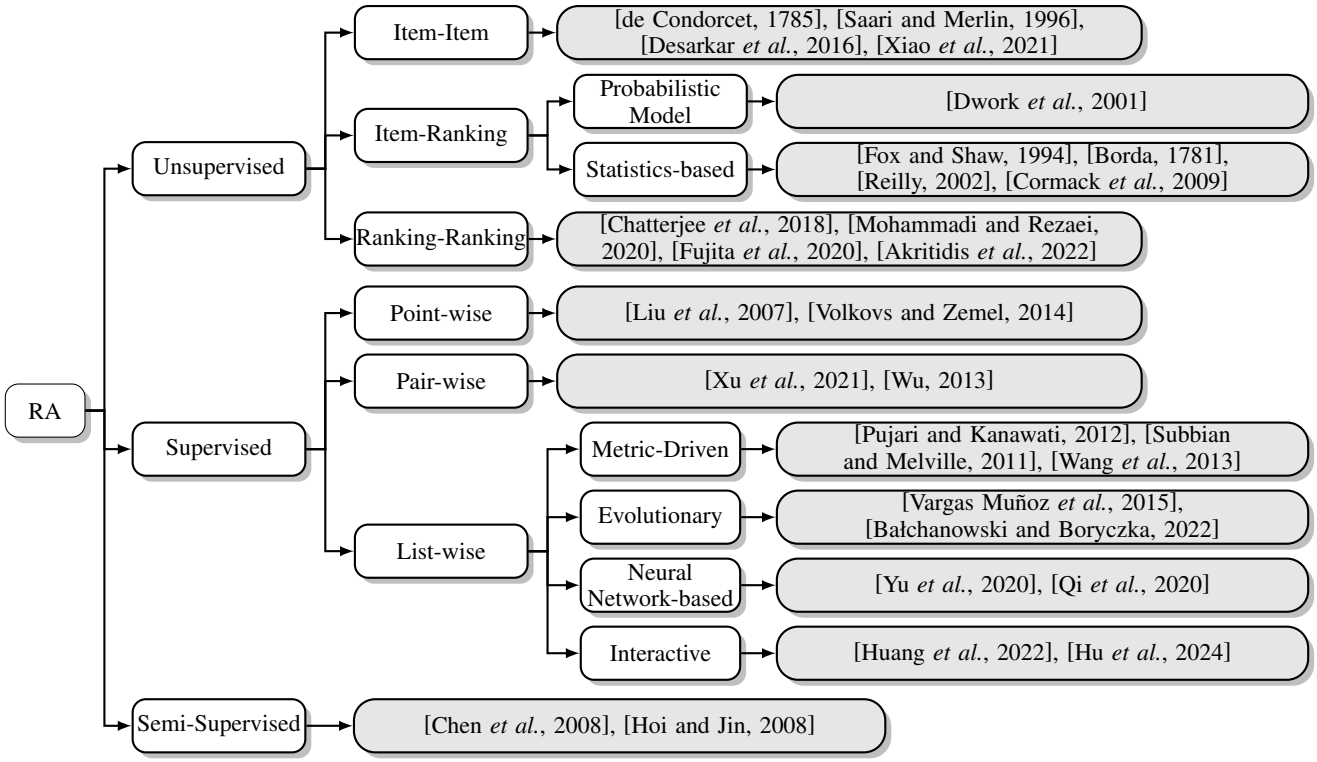

Figure 1: Taxonomy of RA methods.

scores, and items with identical scores are considered ties. Similarly, Outranking [Farah and Vanderpooten, 2007] proposes a comparison method to generate the consensus ranking based on majority opinions, where an item in a higher position in the consensus ranking must have a majority of preferences and, meanwhile, a minority of objections.

Besides, there are several graph-based comparison methods. [Desarkar *et al.*, 2016] treats basic rankings as preference graphs and items as nodes. They calculate node weights through pairwise comparisons of items and then sort the aggregated graphs based on node weights. [Xiao *et al.*, 2021] proposes a competitive graph (CG) method, which ranks a node in the graph to the consensus ranking by calculating its ratio between out- and in-degrees, representing the number of items above and below the node, respectively.

3.2 Item-Ranking Methods

As the most popular class of methods in RA, item-ranking methods utilize positional information of items in basic rankings to produce the consensus ranking. Relevant works can be further subdivided into probabilistic model methods and statistics-based methods.

Probabilistic Model Methods. Probabilistic model methods condense the positional information of items across all basic rankings into a probability model. [Dwork *et al.*, 2001] introduces the Markov chain (MC) to model items' positional information, where each item is modeled as a state. Then a (homogeneous) MC for a basic ranking consists of a set of states $\mathcal{S} = \{u_1, u_2, ..., u_M\}$ and a transition probability matrix $\mathbf{P} \in \mathbb{R}^{M \times M}$. In general, the start state of the system is chosen according to some initial distribution, *e.g.*, the uniform distribution on $\mathcal{S}$. The core idea of [Dwork *et al.*, 2001] is to use the stationary distribution of MC to define the final consensus ranking. Depending on different ways to construct $\mathbf{P}$, there are 4 different RA implementations:

- MC1: if the current state is item u_i, then the next state is chosen uniformly from the multiset of all items u_j that were ranked higher than or equal to u_i by some ranking that ranked u_i, *i.e.*, chose the next state uniformly from the multiset $\mathcal{Q}_i^{\text{MC1}} = \cup_{t=1}^{N}\{u_j : R^t(u_j) \leq R^t(u_i)\}$;
- MC2: if the current state is item u_i, then the next state is chosen by first picking a ranking R^t uniformly from $\mathcal{R}$ containing u_i, then picking an u_j uniformly from the set $\mathcal{Q}_{R^t,i}^{\text{MC2}} = \{u_j : R^t(u_j) \leq R^t(u_i)\}$;
- MC3: if the current state is item u_i, then the next state is chosen as follows: first pick a ranking R^t uniformly from all the $\mathcal{R}$ containing u_i, then uniformly pick an u_j that was ranked by R^t. if $R^t(u_j) < R^t(u_i)$ then go to u_j, else stay in u_i;
- MC4: if the current state is item u_i, then the next state is chosen as follows: first, pick an item u_j uniformly from $\mathcal{S}$. If $R^t(u_j) < R^t(u_i)$ for the majority of the rankings $R^t \in \mathcal{R}$ that ranked both u_i and u_j, then go to u_j, else stay in u_i.

Statistics-based Methods. In general, statistics-based methods utilize a certain statistic to measure the position of the item and use this to get the consensus ranking. [Fox and

CombMIN	$\min_{R^t \in \mathcal{R}} S^t_{\text{Comb}}(u_i)$
CombMAX	$\max_{R^t \in \mathcal{R}} S^t_{\text{Comb}}(u_i)$
CombSUM	$\sum_{R^t \in \mathcal{R}} S^t_{\text{Comb}}(u_i)$
CombANZ	$\frac{1}{\lvert\{R^t \mid R^t \in \mathcal{R}: u_i \in R^t\}\rvert} *$ CombSUM
CombMNZ	CombSUM $* \lvert\{R^t \mid R^t \in \mathcal{R} : u_i \in R^t\}\rvert$

Table 1: Methods of Comb* family [Fox and Shaw, 1994].

Shaw, 1994] proposes Comb* family to aggregate all basic rankings to a consensus one. In the Comb* family, the item u_i in the basic ranking R^t is scored by:

$$S^t_{\text{Comb}}(u_i) = 1 - \frac{R^t(u_i) - 1}{\text{Len}(R^t)} \tag{2}$$

where the $\text{Len}(\cdot)$ returns the number of items in R^t. After computing the scores, the Comb* family traverses all basic rankings and assigns to each item u_i its lowest score (CombMIN), highest score (CombMAX), or arithmetic sum (CombSUM) across all basic rankings as its $S(u_i)$ in the consensus ranking. CombANZ and CombMNZ are CombSUM's variants, in which CombANZ is obtained by dividing CombSUM by the number of rankings that include u_i, while CombMNZ multiplies CombSUM by the number of rankings that include u_i. The Comb* family fusion methods are shown in Table 1.

Fagin [Fagin *et al.*, 2003] proposes the Median method. For each item, they calculate the median position of all basic rankings, and sort all items in descending order based on their median position. BordaCount [Borda, 1781] is the most widely used statistics-based RA method. BordaCount uses the length of ranking to linearly calculate the Borda score of item u_i in basic rankings:

$$S_{\text{Borda}}(u_i) = \sum_{R^t \in \mathcal{R}} \text{Len}(R^t) - R^t(u_i) + 1 \tag{3}$$

The items are then sorted in descending order of their scores to generate the consensus ranking.

Dowdall [Reilly, 2002] designs a reciprocal form score function for RA, which is able to reduce the differences between items when aggregated in the case of a large number of items. The score function of Dowdall is as follows:

$$S_{\text{Dowdall}}(u_i) = \sum_{R^t \in \mathcal{R}} \frac{1}{R^t(u_i)} \tag{4}$$

after scoring items, it arranges the items u_i in descending order based on their scores to generate the consensus ranking. RRF [Cormack *et al.*, 2009] introduces an algorithm that adds a constant k to the denominator of the score function of Dowdall and is able to perform stably with fewer item aggregations. The score function of RRF is as follows:

$$S_{\text{RRF}}(u_i) = \sum_{R^t \in \mathcal{R}} \frac{1}{R^t(u_i) + k} \tag{5}$$

3.3 Ranking-Ranking Methods

The ranking-ranking methods are based on the relationship among various rankings for aggregation. In general, ranking-ranking methods obtain the consensus ranking by a user-defined ranking objective function. iRANK [Wei *et al.*, 2010] propose an unsupervised learning framework that allows two basic rankings to "teach" each other before being combined, to boost the performance of the consensus ranking R^*. ER [Mohammadi and Rezaei, 2020] proposes a method that iteratively solves the consensus ranking based on half-quadratic theory such that the consensus ranking R^* minimizes the distance to each basic ranking. [Chatterjee *et al.*, 2018] introduces a method based on similarity, which computes the similarity matrix between the basic rankings and selects the two rankings with the closest similarity in each round for weighted aggregation, ultimately producing R^*.

HPA [Fujita *et al.*, 2020] proposes a similarity-based method, which firstly generates a pseudo-answer $\bar{R}$ by the simplest mean RA of basic rankings. It then calculates the similarity score between the basic ranking and the pseudo-answer $\bar{R}$. Finally, HPA selects the top-K most similar basic rankings $\mathcal{R}^K$ for aggregation.

$$R^* = \sum_{R_t \in \mathcal{R}^K} \text{sim}(R_t, \bar{R}) R_t \tag{6}$$

where $\text{sim}(\cdot)$ computes the similarity score as the fusion weight. Besides, [Fujita *et al.*, 2020] further introduces a similarity-based method named PostNDCG, which compares the similarity between basis rankings and selects the ranking with the highest similarity to all other rankings as the consensus ranking. Recently, DIBRA [Akritidis *et al.*, 2022] utilizes an innovative approach based on the cosine distance metric to iteratively update weights in basic rankings. It assigns scores to items in each ranking, creating M-dimensional score vectors. DIBRA involves calculating the cosine distance between pairs of basic rankings to identify relevant items, with higher scores indicating greater relevance. By iterating through all rankings, DIBRA determines a consensus ranking.

4 Supervised RA Methods

Given high-quality training data, supervised RA methods generally deliver superior performance. Note that many supervised RA methods generate the consensus ranking R^* through a weighted fusion, thereby aiming to find an optimal fusing weight for $\mathcal{R}$. We define the weight vector as $\boldsymbol{w} = [w_1, w_2, ..., w_N]^\top$, where weight w_t is assigned to R^t. Thus, $R^* = \sum_{R_t \in \mathcal{R}} w_t R^t$.

4.1 Point-wise Methods

[Liu *et al.*, 2007] sets up a general framework for supervised RA, in which learning is formalized as an optimization, which minimizes the disagreements between the consensus ranking R^* and the ground truth that is transferred to a pairwise comparison matrix. Based on this framework, the authors develop supervised versions of BordaCount and MC methods. [Volkovs and Zemel, 2014] incorporates Normalized Discounted Cumulative Gain (NDCG) [Ye *et al.*, 2021]

to construct a loss function and combine it with the Conditional Random Field (CRF) framework. They employ the gradient descent optimization method to train the parameter $\theta = \{\alpha_t, \beta_t^p, \beta_t^n\}_{t=1}^N$ for each basic ranking, which is used to calculate the ranking score of item:

$$S_{\text{CRF}}(u_i) = -\sum_{t=1}^{N} \alpha_t \varphi_t(u_i) - \beta_t^p \sum_{j \neq i} \phi_t(u_i, u_j) + \beta_t^n \sum_{j \neq i} \phi_t(u_j, u_i) \tag{7}$$

where the binary unary potentials $\varphi_t(u_i)$ takes the value 1 only when item u_i is not ranked by R^t and 0 in all other cases. For pairwise potentials ϕ_t, if $R^t(u_i) < R^t(u_j)$, $\phi_t(u_i, u_j)$ is non-zero. Otherwise, $\phi_t(u_j, u_i)$ is non-zero. The weights β_t^p and β_t^n control how much emphasis is given to positive and negative preferences, respectively.

4.2 Pair-wise Methods

[Xu *et al.*, 2021] proposes a deep framework to predict the preference of each item pair while improving the consistency with overall basic rankings as much as possible. [Wu, 2013] proposes a supervised variant of [de Condorcet, 1785], referred to as the weighted Condorcet method, which incorporates a training approach based on Linear Discriminant Analysis (LDA) to optimize the fusing weights $\boldsymbol{w}$ for $\mathcal{R}$. For each item pair (u_i, u_j), according to whether u_i is better than u_j in the ground truth, we can classify all item pairs into two classes. At the same time, each item pair has a feature vector $[h_1, h_2, ..., h_N]^\top$. If $R^t(u_i) < R^t(u_j)$, h_t = 1, otherwise h_t = -1. We want to distinguish the item pairs of the two classes by a linear combination of n features:

$$g(h_1, h_2, \ldots, h_N) = \sum_{i=1}^{N} w_i h_i + w_0 \tag{8}$$

where $\boldsymbol{w}$ is learned by LDA.

4.3 List-wise Methods

List-wise methods focus on the characteristics or attributes of the basic rankings, currently dominating over supervised RA.

Metric-Driven Methods. Using metrics like NDCG, mAP [Ye *et al.*, 2021], and so on, metric-driven methods aim to find the optimal fusing weights $\boldsymbol{w}$ for $\mathcal{R}$. The implementation has two ways: either by using the metric to assess the quality of basic rankings, thereby assigning greater weights to the better ones like wBorda [Pujari and Kanawati, 2012], or by directly optimizing a specific metric to enhance the effectiveness of consensus ranking [Wang *et al.*, 2013].

Evolutionary Methods. Evolutionary algorithms simulate the natural selection, crossover, and mutation process of individuals in the population to find the best individual, which represents the optimal scheme. In [Vargas Muñoz *et al.*, 2015], the individual is represented by a binary tree that defines the aggregate function $f(\mathcal{R})$ with the FFP1 function proposed in [Fan *et al.*, 2004] as the fitness function. AggrankDE [Bałchanowski and Boryczka, 2022] uses the differential evolution algorithm to address the RA problem. It represents an individual as fusing weights $\boldsymbol{w}$ for $\mathcal{R}$, with Average Precision (AP) as the fitness function. Mutation generates a new individual $\boldsymbol{w}'$ by combining three randomly selected individuals $\boldsymbol{w}_j$, $\boldsymbol{w}_h$, and $\boldsymbol{w}_k$: $\boldsymbol{w}' = \boldsymbol{w}_j + \lambda(\boldsymbol{w}_h - \boldsymbol{w}_k)$, where the parameter λ controls the amplification of the differential variation $(\boldsymbol{w}_h - \boldsymbol{w}_k)$.

Neural Network-based Methods. This type of method uses neural networks to optimize aggregation. [Qi *et al.*, 2020] utilizes the Lovasz Bregman (LB) divergence to create a linear structured convex function and a nested structured concave function to aggregate outputs of distributed deep neural network-based models. CSRA [Yu *et al.*, 2020] considers both the quality of basic rankings and the difficulty of each query, and combines the two to train a neural network to predict the fusing weights $\boldsymbol{w}$ for $\mathcal{R}$. At the same time, this method uses the simplest mean fusion [Burges *et al.*, 2011] to construct ground truth without priori information, which, to some extent, overcomes the dependence of supervised methods on labeled data.

Interactive Methods. In the real world, pre-labeled datasets are hardly available, prohibiting fully-supervised RA methods from practical applications. To address this problem, [Huang *et al.*, 2022] utilizes a small amount of supervisory information from users' feedback to supervise the RA method to produce better results. Based on positive (relevant) samples labeled by users, [Huang *et al.*, 2022] proposes two implementations to adjust the fusing weight $\boldsymbol{w}$ for $\mathcal{R}$: IRA_R and IRA_S. IRA_R denotes that the samples are selected based on the position of the item in the ranking, while the IRA_S is the item's score. IRA_R increases weights for rankings that rank positive samples higher, while IRA_S modifies $\boldsymbol{w}$ according to the variance in scores of positive samples, utilizing the standard deviation to reflect the consistency of users' feedback.

5 Semi-supervised RA Methods

As the challenges in the RA field discussed above, semi-supervised methods aim to find a better balance between high-performance and low-cost. [Chen *et al.*, 2008] and [Hoi and Jin, 2008] propose two semi-supervised RA methods, both aim to find an appropriate fusing weight $\boldsymbol{w}$ for $\mathcal{R}$ by using a small amount of available labeled data. [Hoi and Jin, 2008] develops a novel query-dependent solution, in which the $\boldsymbol{w}$ is different based on different characteristics of queries. They use the graph Laplacian matrix to represent the similarity relationships between objects and combine it with a ranking loss function to optimize the $\boldsymbol{w}$.

[Chen *et al.*, 2008] proposes SSRA to learn $\boldsymbol{w}$ by minimizing disagreement between consensus ranking and basic rankings while satisfying preference constraints. The scores given by basic rankings to each item are stored in a $M \times N$ matrix $\mathbf{M}$. Let Σ be the limited set of labeled preference pairs, serving as the constraints. If $(u_i, u_j) \in \Sigma$, then item u_i is ranked higher than item u_j. So, the preference constraints can be

		Datasets													
		Market1501		DukeMTMC-reID		CUHK03 (detected)		CUHK03 (labeled)		MovieLens 1M		NSCLC		WUR 2022	
Method	**Venue**	**R@1**	**mAP**	**R@1**	**mAP**	**R@1**	**mAP**	**R@1**	**mAP**	**R@1**	**mAP'**	**r@400**	**r@800**	**Nor ↓**	**Im-p ↓**
Unsupervised RA															
CombMIN	NIST SP'94	96.08	94.26	91.83	88.88	79.86	80.33	81.71	82.62	38.44	13.04	1.45	7.25	199.66	**1.41**
CombMAX	NIST SP'94	95.64	91.29	89.45	82.62	81.29	78.79	83.79	81.68	38.03	8.86	4.35	11.59	181.72	4.07
CombSUM	NIST SP'94	96.67	92.88	92.06	85.68	83.00	81.06	86.36	84.35	22.50	8.01	8.70	17.39	98.78	1.95
CombANZ	NIST SP'94	94.33	86.40	85.77	76.74	75.50	74.80	81.21	79.21	22.37	7.98	8.70	17.39	98.95	1.96
CombMNZ	NIST SP'94	96.67	89.14	92.15	81.66	83.07	81.78	86.36	84.87	22.40	7.98	8.70	17.39	98.95	1.96
MC1 (top-500)	WWW'01	96.59	93.07	90.80	85.63	83.50	81.30	85.86	84.21	8.73	2.06	4.35	10.14	101.26	2.60
MC2 (top-500)	WWW'01	96.62	93.31	90.93	86.02	83.57	81.65	86.50	84.50	8.63	5.32	5.80	14.49	113.67	2.70
MC3 (top-500)	WWW'01	96.79	93.05	92.24	85.96	83.21	81.23	86.43	84.46	9.35	4.68	5.80	7.25	88.13	2.42
MC4 (top-500)	WWW'01	96.82	93.81	92.46	87.13	84.86	82.75	**88.00**	85.58	9.44	4.52	2.90	8.70	79.74	2.79
BordaCount	SIGIR'01	96.67	92.88	92.06	85.68	83.00	81.06	86.36	84.35	40.20	12.77	10.14	20.29	98.95	1.84
Dowdall	IPSR'02	96.65	93.50	91.20	86.79	84.50	82.30	87.21	85.13	21.18	8.02	10.14	21.74	**78.97**	1.98
Median	SIGMOD'03	96.62	93.37	91.92	85.93	85.29	82.84	87.07	85.09	22.35	6.23	**18.84**	**30.43**	83.01	2.08
RRF	SIGIR'09	96.88	93.41	92.46	86.74	83.64	81.89	86.64	84.85	21.77	8.00	10.14	20.29	90.54	1.98
iRANK	JIST'10	96.50	94.35	92.32	88.17	84.86	83.85	87.00	86.38	22.95	8.14	17.39	24.64	94.86	2.80
Mean	PMLR'11	96.44	94.55	92.24	88.57	84.71	84.40	86.93	86.91	22.50	8.01	17.39	24.64	94.37	2.79
HPA	ECIR'20	96.44	94.74	92.10	**89.09**	81.64	82.23	83.57	84.52	32.68	10.98	15.94	20.29	95.74	2.82
PostNDCG	ECIR'20	96.47	92.09	91.47	84.47	82.50	79.72	83.71	81.92	17.17	6.70	14.49	15.94	149.83	4.29
ER	OMEGA'20	96.64	92.89	92.15	85.68	83.14	81.09	86.36	84.42	39.80	12.70	10.14	20.29	98.82	1.84
CG	JORS'21	96.67	92.88	92.06	85.68	83.00	81.06	86.36	84.35	40.22	12.77	10.14	20.29	98.94	1.84
DIBRA	LSA'22	96.44	94.58	92.37	88.69	84.64	84.39	86.86	86.91	23.54	8.47	17.39	24.64	95.11	2.85
Supervised RA															
wBorda	WWW'12	96.64	92.87	92.15	85.66	83.21	81.17	86.36	84.41	22.50	7.87	-	-	-	-
CRF	CIKM'13	96.64	92.89	92.19	85.67	82.93	81.08	86.36	84.42	50.86	**16.51**	-	-	-	-
CSRA	ICASSP'20	96.44	94.55	92.24	88.81	84.71	84.40	86.93	86.91	-	-	-	-	-	-
AggRankDE	Electronics'22	95.55	93.82	90.44	88.33	75.21	76.98	77.71	79.71	25.68	8.53	-	-	-	-
$IRA_R(1,1)$	BMVC'22	94.86	92.85	90.66	86.52	83.36	82.77	85.14	84.67	22.09	5.84	-	-	-	-
$IRA_R(3,1)$	BMVC'22	97.54	94.70	**93.36**	88.73	**85.50**	84.84	87.79	87.27	44.88	11.89	-	-	-	-
$IRA_S(1,1)$	BMVC'22	96.44	94.56	92.32	88.59	84.79	84.42	86.93	86.91	34.83	9.74	-	-	-	-
$IRA_S(3,1)$	BMVC'22	97.71	94.84	93.63	88.68	85.00	84.39	87.57	87.46	53.53	13.20	-	-	-	-
QI-IRA(1,1)	AAAI'24	96.44	94.55	92.37	88.63	84.79	84.41	87.07	86.93	35.15	10.53	-	-	-	-
QI-IRA(3,1)	AAAI'24	**97.83**	**95.00**	93.31	88.92	**85.50**	**84.86**	**88.00**	**87.54**	**54.37**	16.10	-	-	-	-
Semi-supervised RA															
SSRA	CIKM'08	96.73	92.89	92.10	85.67	83.07	81.14	86.36	84.41	39.55	12.68	-	-	-	-

Table 2: Performance comparison on different datasets. The bold indicates the best and the underline indicates the second best.

given by the inequality:

$$\forall(u_i, u_j) \in \Sigma, \quad \mathbf{M}^{(i)}\boldsymbol{w} - \mathbf{M}^{(j)}\boldsymbol{w} \geq 1. \tag{9}$$

where $\mathbf{M}^{(i)}$ is the i-th row of the matrix $\mathbf{M}$. Then, [Chen *et al.*, 2008] measures the similarity of each basic ranking to all other rankings in $\mathcal{R}$. A ranking that is agreed with more other rankings has higher quality. Let $\boldsymbol{p} = [p_1, p_2, ..., p_N]^\top$ denote the quality vector for R. The optimal $\boldsymbol{w}$ should minimize the distance to $\boldsymbol{p}$ with constraints defined by Eq.(9):

$$\min_{\boldsymbol{w}} \|\boldsymbol{w} - \boldsymbol{p}\|^2 + \lambda_{\boldsymbol{w}} \tag{10}$$

where $\lambda_{\boldsymbol{w}}$ is a regularizer about $\boldsymbol{w}$.

6 Experiments

6.1 Datasets

We conduct benchmarking experiments on 4 popular re-ID datasets, *i.e.*, Market1501 [Zheng *et al.*, 2015], DukeMTMC-reID [Ristani *et al.*, 2016] and CUHK03 detected and labeled [Li *et al.*, 2014], 1 recommendation system dataset, *i.e.*, MovieLens 1M [Harper and Konstan, 2015], 1 bioinformatics dataset, *i.e.*, NSCLC [Wang *et al.*, 2022] and 1 social choice dataset, *i.e.*, World University Ranking in 2022 (WUR 2022) [Feng *et al.*, 2023]. For the quantitative evaluation, we utilize the ranking metrics Rank@1 (R@1 for short) and mAP [Ye *et al.*, 2021] on re-ID datasets, Rank@1 and mAP@10 (mAP' for short) in recommendation system [Oliveira *et al.*, 2020], Recall@400 (r@400 for short) and Recall@800 (r@800 for short) [Wang *et al.*, 2022] in bioinformatics, and Normality (Nor for short) and Impartiality (Im-p for short) in social choice [Feng *et al.*, 2023]. The smaller normality and impartiality are, the better. For all other metrics, the larger the value, the better. All these metrics are in the form of percentages (%), except for normality and impartiality.

6.2 Experiment Setup

All experiments are conducted on a server equipped with 2 Intel Xeon Silver 4215 (2.50GHz) and 4 Nvidia RTX A6000. Specifically, the MC1-4 methods are difficult to test on the full Market1501 and DukeMTMC-reID datasets, so we cut off top-K items from basic rankings and aggregate to be the new MC1-4 (top-K). Besides, there is no labeled data in the NSCLC and the WUR 2022 dataset, so we do not measure supervised and semi-supervised RA methods on them.

Considering the practical situation, we fixed $n = 1$ and only examined the cases $m = 1$ and 3 in interactive methods during the experiment. Due to limitations of space, readers are referred to the project's homepage on GitHub for more

detailed experimental setups, including the generation of basic rankings and the setting of the model parameters.

6.3 Result and Discussion

The result of these RA methods on 7 datasets is shown in Table 2, from which we have the following observations. Overall, unsupervised RA methods can be better applied to datasets in different fields because they do not require supervisory information, and the performance of some unsupervised methods is very close to supervised methods. For supervised methods, there exists at least one that outperforms all other unsupervised methods on the MovieLens 1M and 4 re-ID datasets. The semi-supervised method ranks among the top on the MovieLens 1M and 4 re-ID datasets; despite not being the best, it shows better robustness.

It is remarkable that some unsupervised methods show good performance and robustness on 7 different source datasets, such as Borda, ER, CG, RRF, iRANK, and Mean. The ranking-ranking methods usually achieve a good performance in our experiments, especially for HPA and DIBRA. The HPA method even ranks in the top three of all methods in terms of mAP metrics on both Market1501 and DukeMTMC-reID datasets, with 94.74% and 89.09%, respectively. However, when considering the WUR 2022 dataset, statistics-based methods are superior. Moreover, MC4 obtains the top-2 performance among all unsupervised RA methods on all re-ID datasets for the R@1, especially ranks first on DukeMTMC-reID and CUHK03 (labeled) with 92.46% and 88%, respectively. However, there has been a dramatic decline in the performance of all MC methods on MovieLens 1M. This can be attributed to a lot of noise among basic rankings on MovieLens 1M, and MC methods are not a good choice in such environments.

In supervised RA, QI-IRA(3,1) [Hu *et al.*, 2024] ranks first in the performance of R@1 and mAP for almost all re-ID datasets and MovieLens 1M, except for the mAP@10 in MovieLens 1M, which lost out to the CRF by 0.41% and the R@1 in DukeMTMC-reID, which lost out to the IRA_{S}(3,1) by 0.32%. This can be attributed to the flexible interaction with a small amount of user feedback, which enhances the performance and robustness of the IRA method. Notice that CRF shows very high performance in the MovieLens 1M (R@1: 50.86%, mAP': 16.51%), which can be mainly attributed to the CRF framework's ability to utilize low-quality basic rankings, meaning it can offer a certain level of reverse support for the preference relations they endorse.

Furthermore, we can observe that the semi-supervised method consistently ranks at the top among all RA methods, exhibits good robustness on 4 re-ID datasets, and achieves 39.55% R@1 on MovieLens 1M. This suggests that supervised RA methods are a preferable choice when there is a significant performance difference among basic rankings and labeled data is available. However, the semi-supervised method can be considered as an alternative when the cost budget is insufficient.

7 Open Questions

7.1 Dilemma of High-Performance and Low-Cost

The earliest exploration to address the dilemma of high-performance and low-cost is semi-supervised RA [Hoi and Jin, 2008] [Chen *et al.*, 2008]. While these methods significantly reduce data labeling costs, they still incur considerable data collection costs. Moreover, static training data are not flexible enough to adapt to changing real-world scenarios. To address this problem, [Huang *et al.*, 2022] proposes the interactive RA. It utilizes a small amount of users' online feedback to guide the RA process and achieves the dual advantages of higher performance than unsupervised RA at a low cost. On this basis, [Hu *et al.*, 2024] introduces quantum theory to provide a theoretical explanation for the generation and aggregation of multiple rankings, and raises a new quantum-inspired interactive RA method QI-IRA. However, the current IRA method selects samples solely depending on users' subjective judgment, which may result in low-efficient interactions for inexperienced users. The active learning (AL) technique has the ability to objectively recommend valuable feedback candidates, which can effectively address the above defects. Therefore, how to introduce AL into interactive RA is a promising research direction for interactive RA.

7.2 New Application of RA Technique

With the widespread applications of Deep Neural Networks (DNNs), adversarial defense for DNNs has become increasingly important. Most successful defense strategies adopt adversarial training or random input transformations that typically require retraining or finetuning the model to achieve reasonable performance. [Tiwari *et al.*, 2022] finds that employing the RA technique to combine the network hidden layer outputs shows unexpected robustness to adversarial attacks. However, when evaluating the robustness of the model, the above work always assumes that attacks only target fixed output ports, which results in an overestimation of the robustness of the model. To address the above problem, if we can effectively eliminate the interference of structural mismatch between attack and defense, we may better utilize the RA technique to defend DNN so as to produce more robust and trusted prediction results.

8 Conclusion

In this paper, we present a comprehensive review of classical methods and recent advances in RA study. Specifically, we first elaborate on three key challenges facing RA research, *i.e.*, improving performance, reducing cost, and striking a balance between them. On this basis, we analyze and discuss representative algorithms in unsupervised, supervised and semi-supervised RA. For each category, we summarize their common ideas and comment on their strengths and weaknesses. To compare the performance of different RA methods, we conduct the largest scale of comparative evaluation to date for 27 RA methods on 7 public datasets from re-ID, recommendation system, bioinformatics and social choice. Ultimately, we discuss open questions in the RA study and give our suggestions for future prospects.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62372339). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

Contribution Statement

Siyi Wang, Qi Deng, and Shiwei Feng contributed equally to this work.

References

[Akritidis *et al.*, 2022] Leonidas Akritidis, Athanasios Fevgas, Panayiotis Bozanis, and Yannis Manolopoulos. An unsupervised distance-based model for weighted rank aggregation with list pruning. *Expert Systems with Applications*, 202:117435, 2022.

[Bałchanowski and Boryczka, 2022] Michał Bałchanowski and Urszula Boryczka. Aggregation of rankings using metaheuristics in recommendation systems. *Electronics*, 11(3):369, 2022.

[Borda, 1781] J-C de Borda. Mémoire sur les élections au scrutin: Histoire de l'académie royale des sciences. *Paris, France*, 12, 1781.

[Burges *et al.*, 2011] Christopher Burges, Krysta Svore, Paul Bennett, Andrzej Pastusiak, and Qiang Wu. Learning to rank using an ensemble of lambda-gradient models. In *Proceedings of the learning to rank Challenge*, pages 25–35. PMLR, 2011.

[Chatterjee *et al.*, 2018] Sujoy Chatterjee, Anirban Mukhopadhyay, and Malay Bhattacharyya. A weighted rank aggregation approach towards crowd opinion analysis. *Knowledge-Based Systems*, 149:47–60, 2018.

[Chen *et al.*, 2008] Shouchun Chen, Fei Wang, Yaangqiu Song, and Changshui Zhang. Semi-supervised ranking aggregation. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 1427–1428, 2008.

[Cormack *et al.*, 2009] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759, 2009.

[de Condorcet, 1785] Marquis de Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions, rendues à la pluralité des voix/— c par m. le marquis de condorcet... 1785.

[Desarkar *et al.*, 2016] Maunendra Sankar Desarkar, Sudeshna Sarkar, and Pabitra Mitra. Preference relations based unsupervised rank aggregation for metasearch. *Expert Systems with Applications*, 49:86–98, 2016.

[Dwork *et al.*, 2001] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.

[Fagin *et al.*, 2003] Ronald Fagin, Ravi Kumar, and Dandapani Sivakumar. Efficient similarity search and classification via rank aggregation. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 301–312, 2003.

[Fan *et al.*, 2004] Weiguo Fan, Edward A Fox, Praveen Pathak, and Harris Wu. The effects of fitness functions on genetic programming-based ranking discovery for web search. *Journal of the American Society for Information Science and Technology*, 55(7):628–636, 2004.

[Farah and Vanderpooten, 2007] Mohamed Farah and Daniel Vanderpooten. An outranking approach for rank aggregation in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 591–598, 2007.

[Feng *et al.*, 2023] Shiwei Feng, Qi Deng, Siyi Wang, Lin Song, and Chao Liang. An experimental study of unsupervised rank aggregation methods in world university rankings. In *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)*, pages 1–8, 2023.

[Fox and Shaw, 1994] Edward Fox and Joseph Shaw. Combination of multiple searches. *NIST special publication SP*, pages 243–243, 1994.

[Fujita *et al.*, 2020] Soichiro Fujita, Hayato Kobayashi, and Manabu Okumura. Unsupervised ensemble of ranking models for news comments using pseudo answers. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 133–140. Springer, 2020.

[Harper and Konstan, 2015] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

[Hoi and Jin, 2008] Steven CH Hoi and Rong Jin. Semi-supervised ensemble ranking. In *AAAI*, volume 8, pages 634–639, 2008.

[Hu *et al.*, 2024] Chunyu Hu, Hong Zhang, Chao Liang, and Hao Huang. Qi-ira: Quantum-inspired interactive ranking aggregation for person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2202–2210, 2024.

[Huang *et al.*, 2022] Ji Huang, Chao Liang, Yue Zhang, Zhongyuan Wang, and Chunjie Zhang. Ranking aggregation with interactive feedback for collaborative person re-identification. 2022.

[Li *et al.*, 2014] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[Li *et al.*, 2019] Xue Li, Xinlei Wang, and Guanghua Xiao. A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications. *Briefings in bioinformatics*, 20(1):178–189, 2019.

[Lin, 2010] Shili Lin. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):555–570, 2010.

[Liu *et al.*, 2007] Yu-Ting Liu, Tie-Yan Liu, Tao Qin, Zhi-Ming Ma, and Hang Li. Supervised rank aggregation. In *Proceedings of the 16th international conference on World Wide Web*, pages 481–490, 2007.

[Mohammadi and Rezaei, 2020] Majid Mohammadi and Jafar Rezaei. Ensemble ranking: Aggregation of rankings produced by different multi-criteria decision-making methods. *Omega*, 96:102254, 2020.

[Oliveira *et al.*, 2020] Samuel EL Oliveira, Victor Diniz, Anisio Lacerda, Luiz Merschmanm, and Gisele L Pappa. Is rank aggregation effective in recommender systems? an experimental analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(2):1–26, 2020.

[Pujari and Kanawati, 2012] Manisha Pujari and Rushed Kanawati. Supervised rank aggregation approach for link prediction in complex networks. In *Proceedings of the 21st International Conference on World Wide Web*, pages 1189–1196, Lyon France, April 2012. ACM.

[Qi *et al.*, 2020] Jun Qi, Chao-Han Huck Yang, and Javier Tejedor. Submodular rank aggregation on score-based permutations for distributed automatic speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3517–3521. IEEE, 2020.

[Reilly, 2002] Benjamin Reilly. Social choice in the south seas: Electoral innovation and the borda count in the pacific island countries. *International Political Science Review*, 23(4):355–372, 2002.

[Renda and Straccia, 2003] M Elena Renda and Umberto Straccia. Web metasearch: rank vs. score based rank aggregation methods. In *Proceedings of the 2003 ACM symposium on Applied computing*, pages 841–846, 2003.

[Ristani *et al.*, 2016] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[Saari and Merlin, 1996] Donald G Saari and Vincent R Merlin. The copeland method: I.: Relationships and the dictionary. *Economic theory*, 8:51–76, 1996.

[Subbian and Melville, 2011] Karthik Subbian and Prem Melville. Supervised rank aggregation for predicting influence in networks. *arXiv preprint arXiv:1108.4801*, 2011.

[Tiwari *et al.*, 2022] Lokender Tiwari, Anish Madan, Saket Anand, and Subhashis Banerjee. Regroup: Rank-aggregating ensemble of generative classifiers for robust predictions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2595–2604, 2022.

[Vargas Muñoz *et al.*, 2015] Javier Alvaro Vargas Muñoz, Ricardo da Silva Torres, and Marcos André Gonçalves. A soft computing approach for learning to aggregate rankings. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 83–92, 2015.

[Volkovs and Zemel, 2014] Maksims N Volkovs and Richard S Zemel. New learning methods for supervised and unsupervised preference aggregation. *The Journal of Machine Learning Research*, 15(1):1135–1176, 2014.

[Wang *et al.*, 2013] Yang Wang, Yalou Huang, Xiaodong Pang, Min Lu, Maoqiang Xie, and Jie Liu. Supervised rank aggregation based on query similarity for document retrieval. *Soft Computing*, 17:421–429, 2013.

[Wang *et al.*, 2022] Bo Wang, Andy Law, Tim Regan, Nicholas Parkinson, Joby Cole, Clark D Russell, David H Dockrell, Michael U Gutmann, and J Kenneth Baillie. Systematic comparison of ranking aggregation methods for gene lists in experimental results. *Bioinformatics*, 38(21):4927–4933, 2022.

[Wei *et al.*, 2010] Furu Wei, Wenjie Li, and Shixia Liu. irank: A rank-learn-combine framework for unsupervised ensemble ranking. *Journal of the American Society for Information Science and Technology*, 61(6):1232–1243, 2010.

[Wu, 2013] Shengli Wu. The weighted condorcet fusion in information retrieval. *Information processing & management*, 49(1):108–122, 2013.

[Xiao *et al.*, 2021] Yu Xiao, Hong-Zhong Deng, Xin Lu, and Jun Wu. Graph-based rank aggregation method for high-dimensional and partial rankings. *Journal of the Operational Research Society*, 72(1):227–236, 2021.

[Xu *et al.*, 2021] Qianqian Xu, Zhiyong Yang, Zuyao Chen, Yangbangyan Jiang, Xiaochun Cao, Yuan Yao, and Qingming Huang. Deep partial rank aggregation for personalized attributes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 678–688, 2021.

[Ye *et al.*, 2021] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021.

[Yu *et al.*, 2020] Yinxue Yu, Chao Liang, Weijian Ruan, and Longxiang Jiang. Crowdsourcing-based ranking aggregation for person re-identification. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1933–1937. IEEE, 2020.

[Zhang *et al.*, 2021] Yang Zhang, Yu Xiao, Jun Wu, and Xin Lu. Comprehensive world university ranking based on ranking aggregation. *Computational Statistics*, 36:1139–1152, 2021.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.