PhD Dissertation

International Doctorate School in Information and Communication Technologies

DISI - University of Trento

Ranking Aggregation Based on Belief Function Theory

Andrea Argentini

Advisor: Prof. Enrico Blanzieri

Università degli Studi di Trento

March 2012

Abstract

The ranking aggregation problem is that to establishing a new aggregate ranking given a set of rankings of a finite set of items. This problem is met in various applications, such as the combination of user preferences, the combination of lists of documents retrieved by search engines and the combination of ranked gene lists. In the literature, the ranking aggregation problem has been solved as an optimization of some distance between the rankings overlooking the existence of a true ranking. In this thesis we address the ranking aggregation problem assuming the existence of a true ranking on the set of items: the goal is to estimate an unknown, true ranking given a set of input rankings provided by experts with different approximation quality. We propose a novel solution called Belief Ranking Estimator (BRE) that takes into account two aspects still unexplored in ranking combination: the approximation quality of the experts and for the first time the uncertainty related to each item position in the ranking. BRE estimates in an unsupervised way the true ranking given a set of rankings that are diverse quality estimations of the unknown true ranking. The uncertainty on the items's position in each ranking is modeled within the Belief Function Theory framework, that allows for the combination of subjective knowledge in a non Bayesian way. This innovative application of belief functions to rankings, allows us to encode different sources of a priori knowledge about the correctness of the ranking positions and also to weigh the reliability of the experts involved in the combination. We assessed the performance of our solution on synthetic and real data against state-of-the-art methods. The tests comprise the aggregation of total and partial rankings in different empirical settings aimed at representing the different quality of the input rankings with respect to the true ranking. The results show that BRE provides an effective solution when the input rankings are heterogeneous in terms of approximation quality with respect to the unknown true ranking.

Keywords
[Ranking Aggregation, Belief function Theory, Dempster-Shafer Theory]

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

Ranking aggregation is a relevant problem that is faced in several application contexts such as marketing research, psychology and meta-search. In its general form the problem is stated as follows: given a set of experts or judges providing an ordered list from a set of items, the goal is to find a list that best represents the wholeset of input rankings according to some measure. A practical context that has given large in impulse to ranking aggregation solutions is the so-called meta-search where the results of several search engines have to be combined to produce a consensus answer. Also in Bioinformatics, the aggregation of rankings emerges from the need to integrate different biological data related to the same question to be investigated. Although ranking aggregation is an optimization problem based on distance, it been shown that the solution of the Kemeney optimal aggregator (with Kendall's distance) is an NP-hard problem even with justfour rankings. This computational limitation has led to several solutions to alleviate the computational burden of the problem. The ranking aggregation methods can be divided into two groups, one which comprises stochastic optimization methods and the other that includes heuristic methods. The methods in the first group try to find the best aggregated ranking using an optimization method, whereas the heuristic methods approximate the solution by means of heuristics. The rankings can be divided into three kinds: total rankings, partial and top-k rankings (lists). The difficulty of the aggregation is greater in the case of partial or top-k lists since the lists have different lengths and share also disjoint sets of items.
We point out that in the formulation of the ranking aggregation problem the quality of input rankings and the presence of a true ranking is overlooked. Few works in literature have arisen this fact, highlighting that the ultimate goal is to find an aggregate ranking closer to the true ranking. In this work we tackle the ranking aggregation problem introducing the true ranking in its formulation. The goal is to find a satisfying estimatate of the unknown true ranking given a set of input rankings provided by experts with different degree of approximation quality. We claim that this the case for rankings provided by bioinformatic experts because of the underlying physical reality of the unknown biological phenomenon at hand.

To solve the ranking aggregation problem as discussed above we propose a solution based on the Belief Function Theory. The Belief Function theory, also called Dempster-Shafer theory, is a powerful framework for reasoning with imprecise and uncertain data,

allowing the modeling of subjective knowledge in a non Bayesian way. Given a frame of possible hypotheses or items, the framework allows us to assign a quantitative measure of the expert evidence on the whole power set of the frame. This leads to model various levels of knowledge of the experts from complete knowledge down to total ignorance. Several combination rules and conditioning operations are defined in the framework, to update and combine the beliefs of the experts. Moreover, the framework extends both the usual set theory operations (union, intersection) and probability theory (conditioning, marginalization). The Belief Function Theory has been applied to machine learning problems such as classification, clustering and combination of classifiers.
In this thesis we propose and evaluate a novel algorithm, called Belief Ranking Estimator (BRE) that estimates the true ranking given a set of rankings. Through the use of Belief Functions we model the correctness of the items ranked from the point of view of each expert (ranking): we then combine all the experts views taking into account the reliability of the rankings involved. The reliability of the input rankings is assessed by computing the distance at the rankings to a true-rank estimator. As the true-rank estimator we can to use the output ranking of any aggregation method.

The novelty of our solution lies in a new formulation of the ranking aggregation problem that takes into account the quality of the input rankings with respect to the true ranking. A second aspect is the modeling of the correctness of the rank in order to manage the uncertainty of the items ranked, using Belief Functions. To the best of our knowledge this framework has never been applied to ranking aggregation before. Moreover, this is the first approach in the ranking aggregation literature that deals with the uncertainty of ranked items. One of the aadvantages is that our approach alloes to model different pieces of *a priori* knowledge about the experts involved (such as the correctness of the positions of a subset of the items with respect to the others) into the aggregation step.
Several possible extensions of the algorithm may be proposed. We have focused our efforts on the empirical results of our method instead of pursuing a more theoretical analysis. The performance of our solutions has been compared against state-of-the-art methods, on both synthetics and real data. With respect to total rankings, we have evaluated the performance of BRE using different true-rank estimators. Moreover, the role of the weights in the algorithm has been deeply investigated. A novel algorithm, called Quality Belief Ranking Estimator (QBRE), for the approximation of the quality of the input rankings has been proposed and evaluated on total rankings. Due to the lack of real data containing total rankings with an available true ranking, we have developed a rigorous experimental setting based on synthetic data. On the partial/top-k rankings we have investigated the performance of BRE both on the synthetic data and on real data. On synthetics data we have tested BRE on three cases of partial rankings that meet different hypotheses on the quality of the experts. Finally, BRE has been evaluated on LETOR, that is a collection of datasets related to the meta-search problem.

The empirical results of BRE on total rankings have showed that BRE outperforms the competitor methods in the cases where the input rankings have heterogeneous quality with respect to the true rankings. The results have showed how the use of the weights is

one of key points in our solution. Weights that are good estimation of the quality of the input rankings increase considerably the performance of BRE. The evaluation of BRE on partial/top-k rankings has highlighted some difficulties to outperform the competitors. As for partial rankings, we have also showed how BRE can encode different *a priori* information such as different belief assignments.
This thesis in structured in the following chapters. Chapter 1 is devoted to the introduction of the ranking aggregation problem and of the main state-of-the-art. A basic introduction of the concepts and the operators of Belief Function Theory is also provided. In Chapter 2, our method, the Belief Ranking Estimator, is introduced and explained in all its components. Some numerical examples are also provided. A version of BRE for the estimation of the quality of the input rankings is also presented. The rest of the chapter is devoted to the experimental evaluations of all the methods proposed for the total rankings. In Chapter 3, BRE is evaluated on the aggregation of partial/top-k rankings. The modifications of BRE to process partial rankings are also discussed. The remainder of the chapter is devoted to experimental results of the application of BRE to synthetic data and the LETOR datasets. In Chapter 4 we draw the final conclusions of this work and highlight possible future directions of works.

Chapter 2

Ranking Aggregation and Belief Functions: An Introduction

2.1 The Ranking Aggregation Problem and its Application

Ranking aggregation is a well-studied problem, that arises in different areas such as psychology, market advertisement research, combination of experts in Information Retrieval or in Bioinformatics. Rankings speaking generally, are ordered sets of items where the order can be provided in several ways: for example, from the subjective preference of users or from the output numeric score of an algorithm (e.g. a classifier). The problem of ranking aggregation concerns the combination of several rankings in order to obtain a final ranking that satisfies specific criteria.

The ranking is a really simple structure that encodes in a intrinsic way some of the knowledge that brings on expert or an user to generate an ordered list. On the other hand, the hidden information that has generated the ranking is really difficult to know and use in the aggregation step. For this reason ranking aggregation methods are based only on the information of the rankings: the items and their position. Rankings are defined as total if they are permutations of the set of items. In many situations, rankings are partial: not all the lists contain the same items. A particular case of partial lists are top-k lists, where only the first k items are included in the rankings. However, in most of the real applications (for example document meta-search) partial/top-k lists are the only rankings available, and the partiality of the rankings increases the difficulties of the aggregation task.

The wide diffusion of rankings derives from the easy way to obtain them from existing data or experimental outcomes. In user-profiling, user preferences are obtained via interviews: in data mining , rankings are easily generated by classifiers outcomes. The easiness to obtain rankings from data and their simple way to describe the expert view underlying the data, are well suited for a scenario of integration where experts provide rankings from different kind of data related to the same problem to investigate. This scenario is frequently encountered in Bioinformatics, where data from different omics (proteomics, metabolomics) and measured on different (and not directly comparable) technologies are integrated to find a list of common genes related to a particular disease or biological

condition [1]. Other Bioinformatics tasks where ranking aggregation has been applied are the combination of miRNA target prediction and the combination of different gene expression microarray studies [2]. Ranking aggregation has also been widely applied to the meta-search problem, which concerns the combination of the answers of several web search engines [3][4][5].
Statistics has given a notable contribution to the study of distances between rankings [6] and of probabilistic models able to deal with the combination of rankings, such as Thurstone's model[7][8], Luce's model [9][10], and Mallow's model with its extension to partial data [11][12][13]. The solutions provided for the ranking aggregation problem are basically of two categories: Stochastic optimization and heuristic methods. In the first category, optimization techniques, such as cross-entropy Monte Carlo [2][14] and genetics algorithms [15], are applied to obtain the optimal aggregation for a given distance measure. Heuristic methods are simpler methods that find a solution based on their own criteria, such as the Borda Count's methods that include the mean and the median of the rankings [16][4], the Markov Chain [4] and MEDrank [17]. Moreover unsupervised and supervised algorithms based on generative models for rankings have been applied to ranking aggregation on meta-search problems [5][18][19].

In the following section we briefly define formally the rankings and the major distances used in rankings aggregation. Sec. 2.3 is devoted to the presentation of the state-of-the-art solutions for aggregating rankings. Finally, we introduce a slightly different point of view on ranking aggregation based on by the presence a true ranking underlying the real situation.

2.2 Definition of Rankings and Distance between Rankings

Before describing the ranking aggregation problem and the different types of solutions proposed in literature, we introduce and define the rankings and the most used distances between them. We begin with total rankings, ranking is a permutation of a set of objects. Let $X = \{x_1, \ldots, x_n\}$ be a set of items to be ranked by an expert opinion. We refer indistinctly to the objects in X as elements or items. We denote as $\tau = (\tau(1), \ldots, \tau(n))$ a ranking associated to X, where $\tau(i)$ is the rank associated with the item x_i. Each expert knows all the elements in X, and it provides an ordering of its elements. We denote as R_j the j-th expert involved in the ranking, so for each expert we have a corresponding ranking $\tau^{R_j} = (\tau^{R_j}(1), \ldots, \tau^{R_j}(n))$. To simplify the notation, each ranking is denoted by τ^j for all $j \in N$ where N is the number of experts and consequently of rankings. The rank values associated to the most important element can be either 1 or n, without any loss of generality in the permutation case. The notation $|\tau^j|$ means the length of the ranking and in case of total rankings for all $j \in N \quad n = |\tau^j|$. We point out that for sets specified with $\{\}$ the ordering of the elements is arbitrary, whereas when using $()$, a specific order with respect to the rank of the items is given.
The most popular distances between rankings are Spearman's footrule and the Kendall's distance. The Spearman distance, is the sum of the absolute differences between the rank

values of the rankings. We define the footrule distance as follows:

$$F(\tau, \sigma) = \sum_{i=1}^{n} |\tau(i) - \sigma(i)| \tag{2.1}$$

where τ and σ are total rankings. As indicated in [6], the F distance can be normalized by dividing by the maximum value $\frac{n^2}{2}$, so that an F value equal to 1 means totally different rankings and 0 means identical rankings. The F distance is computable in linear time.

The Kendall distance [20] compares rankings, counting the pairwise disagreements between two rankings. Formally, Kendall's distance [21] between two total rankings τ and σ is defined as:

$$K(\tau, \sigma) = \sum_{\{i,j\} \in P} K^{*}_{i,j}(\tau, \sigma) \tag{2.2}$$

where P is the set of the unordered pairs of distinct items in X and $K^{*}_{i,j}(\tau, \sigma)$ is defined as:

$$K^{*}_{i,j}(\tau, \sigma) = \begin{cases} 0 & \text{if } x_i,\ x_j \text{ are in the same order in } \tau \text{ and } \sigma \\ 1, & \text{if } x_i, x_j \text{ are in the inverse order in } \tau \text{ and } \sigma \end{cases} \tag{2.3}$$

Kendall's distance can be normalize by dividing by its maximum value $\binom{n}{2}$ [6]. It turns out to be the number of adjacent transpositions needed to transform one ranking into the other and it can be computed in $n \log n$ time.
Another measure to evaluate correlation between two rankings is Spearman's rank correlation coefficent [20][22]. Given two total rankings τ and σ, Spearman's correlation coefficent, denoted by ρ, is defined as :

$$\rho(\sigma, \tau) = 1 - \frac{6 \sum_{i=1}^{n} (\pi(i) - \sigma(i))^2}{n(n^2 - 1)} \tag{2.4}$$

ρ is defined as the Pearson correlation between two ranked variables, namely rankings. ρ returns values in the interval $[-1, 1]$. $\rho = 1$ means a total positive correlation between the rankings insted $\rho = -1$ means a total negative correlation between the input rankings.

If the items present in the rankings are not the same, two different situations can arise: Partial rankings and top-k rankings. Partial rankings, also referred to as partial lists, occur when the rankings are induced by a total ordering over a subset of X. We denote as $\tau^j = (\tau^j(1), \ldots, \tau^j(i), \ldots, \tau^j(l_j))$ the partial ranking of length $l_j = |\tau^j|$, where C_j denotes the set of items in the j-th ranking and $x_i \in C_j$, $C_j \subset X$. This situation is really common in real problems, for example in meta-search applications, search engines return a list of documents for a query that contains a far fewer of documents than the number of all web pages. In this case we cannot make any assumption for the documents not ranked by the expert, since the relation between the subset of items and the universe set U is unknown. Top-k rankings are a special case of partial rankings for which only the top k elements are included in the output rankings of the experts. The top-k rankings are still denoted by $\tau^j = (\tau^j(1), \ldots, \tau^j(i), \ldots, \tau^j(l_j))$, with length of $k_j = l_j = |\tau^j|$ and $x_i \in C_j$, $C_j \subset X$. Assuming that the experts know the same set of items, it is reasonable to assume that the unranked items in a top-k list can be placed below, with the same rank values. A detailed description of the hypothesis evaluated on partial ranking/top-k lists is discussed in Chapter 4

To deal with the comparison between partial rankings and total rankings, a suitable generalizations of the distances on total rankings are to be defined. We denote as $\tau_{|S}$ the projection of a ranking τ with respect to a subset S this produces a new ranking that contains only the element in S and maintains the order of τ.
Induced Footrule Distance Given $\tau^1, \dots, \tau^N$ partial rankings, let U be the union of the elements in $\tau^1, \dots, \tau^N$ and σ a total ranking w.r.t U. The induced footrule distance [4] between σ and a partial list τ^j is:

$$i.F = F(\sigma_{|\tau^j}, \tau^j) \tag{2.5}$$

where the $\sigma_{|\tau^j}$ is the projection of the total ranking σ on the elements of the partial ranking τ^j. In this case the result $\sigma_{|\tau^i}$ is re-ranked in order to compute the F distance. The normalization of the $i.SF$ is done dividing by $\frac{|\tau^j|^2}{2}$, since it is an F distance computed on the same items. In case of N partial rankings the induced footrule distance is:

$$i.F(\sigma, \tau^1, \dots, \tau^N) = \frac{\displaystyle\sum_{j=1}^{N} F(\sigma_{|\tau^j}, \tau^j)}{N}$$

Scaled Footrule Distance The scaled footrule distance [4] is defined as follows:

$$s.F(\sigma, \tau) = \sum_{i \in C_\tau} |\sigma(i)/|\sigma| - \tau(i)/|\tau|| \tag{2.6}$$

where σ, τ are respectively the total ranking and the partial ranking and C_τ is the set of elements ranked in τ. The main difference between the scaled one with respect to the induced version, is the weighting of the distance based on the size of the rankings. We have normalized the $s.F$ distance dividing by $|\tau|/2$ as suggested in [4]. We presented only the distances related to the Spearman footrule distance since we deal only with this distance in this work. For Kendall distance versions for partial and total rankings have also been proposed [21]. An easy way to manage and compare top-k rankings is to transform the top-k rankings in a sort of total rankings (so-called augmented rankings [5]) where in each list the unranked items are placed at position $k + 1$ and finally apply the usual distances.

2.3 Ranking Aggregation Methods

In this section we provide a discussion of state-of-the-art methods and a formal definition of the ranking aggregation problem.

The goal in the ranking aggregation problem is to find a ranking that minimizes the distance from the input rankings, given a ranking distance. One desirable criterion to satisfy is the Kemeny optimal aggregation. Given a set of total rankings $\tau^1, \dots, \tau^j, \dots, \tau^N$

on the set of items X, the goal is to find the underlying order on X, τ. Suppose that the input rankings τ^j are noisy versions of τ, obtained by swapping two elements of τ with a probability $p < 1/2$. The maximum likelihood estimate of τ using the Kendall distance K is [23]:

$$\tau^* = \arg\min_{\tau} \frac{1}{N} \sum_{j=1}^{N} K(\tau, \tau^j) \tag{2.7}$$

The estimate τ^* is referred as to Kemeny optimal aggregation [4]. Dwork *et al.* have shown that the computation of the Kemeny optimal aggregation is a NP-hard problem even when the number of rankings is four [4]. In order to solve the Kemeny optimal aggregation-problem, *stochastic search solutions* and *heuristic solutions* has been proposed.

The Kemeny optimal aggregation (Eq. 2.7) can be approximated via the Spearman footrule distance [4]. This leads to the Footrule optimal aggregation where the above optimal aggregation criteria is based on the Spearman footrule distance. The Footrule optimal aggregation for total rankings is computable in polynomial time by reduction to the computation of the minimum cost of matching on weighted bipartite graph [4]. Let $\tau^1, \ldots, \tau^j, \ldots, \tau^N$ be N total rankings over a universe set X ($n = |X|$). We define a weighted bipartite graph (X,P,W), where $X = \{x_i, \ldots, x_n\}$ is the set of the items to be ranked and $P = \{1, \ldots, n\}$ contains the n possible positions $p \in P$. Each node of X is connected to all the possible positions. The weight for each edge $W(x_i, p)$ is set to the total footrule distance of the rankings that rank item x_i item at position p. This corresponds to $W(x_i, p) = \sum_{j=1}^{N} F(\tau^j(i), p)$. The output is the permutation over the set X that results by the minimum cost of perfect matching in the bipartite graph.
In the case of partial lists, finding the Footrule optimal aggregation is an NP-hard problem, Dwork *et al.* suggest to solve the problem (as in the case of total rankings) as the minimum cost of a bipartite graph in which the weights assigned are based on the scaled footrule distance [4].

Stochastic Optmimization

In order to efficiently explore the combinatorial solution space to find the τ^* from Eq. 2.7, a stochastic method called, cross-entropy Monte Carlo (CEMC) has been proposed [2]. Cross-entropy Monte Carlo is a quite complex stochastic search, that at each step chooses the parameter that minimizes a cross-entropy measure from the actual probability matrix and a sample drawn from the same distribution. Starting from the initial probability matrix, the goal is to update step by step the matrix in order to place more probability on the items that are neighborhood of the best solution. For a more exhaustive description of the method we refer to the original proposal [2]. The CEMC solution has been evaluated on several bioinformatic tasks such as the integration of micro-RNA target prediction and microarray data showing good results with respect to the Markov

Chain solutions [2]. It has also been applied in another bioinformatics work where the task was to find the best clustering algorithms across different evaluation measures space [14]. In that work it has been introduced a weighted formulation of both the footrule and Kendall distances. The weights used are the quantitative outputs of the methods (scores), in order to penalize the difference of the rank of each item. The same authors have implemented the CEMC algorithm in a R package called RankAggr [15]. In the same work an optimization algorithm based on genetic algorithms is also proposed.

Heuristic Solutions for Ranking Aggregation

Methods which provide approximate solutions without optimizing any cost function, are classified as heuristic. This category include the Borda Count [16][4], the MEDrank algorithms, and also other simple heuristics such as the median and the mean of the rankings that can be generalized by the Borda Count.
Borda Count is a really simple method that can include different aggregation functions. Borda Count assigns to each item x_i a score $B(i)$ corresponding to the position in which the item appears in a specific ranking. For each item all the scores are summed up for all the rankings and finally the items are ranked by their total score. Given N total rankings $\tau^1, \dots, \tau^j, \dots, \tau^N$ for each item $x_i \in X$ and for each ranking τ^j Borda Count assigns a score $B^j(i)$ equal to the number of items placed below x_i in τ^j. Let $B_i = f(B^1(1), \dots, B^j(i), \dots, B^N(n))$ an aggregate function of the Borda scores where $i \in \{1, \dots, n,\}$ the final rankings is obtained sorting the B_i score. The most used aggregate functions [1][4] are:

- the median $f(B^1, \dots, B^N) = median(B^1, \dots, B^N)$
- the geometric mean $f(B^1, \dots, B^N) = \left(\prod_{l=1}^{N} B^l \right)^{1/N}$
- the p-norm $f(B^1, \dots, B^N) = \sum_{l=1}^{N} (B^l)^p$

We point out that the arithmetic mean is a special cases of the p-norm when $p = 1$ and $B^j(i) = \tau^l(i)$. In the case of partial rankings the Borda Count works as in the total ranking case, the only difference is that an equal score is assigned to the unranked items for in each partial ranking.

Another heuristic proposed to solve the Kemeny optimal aggregator is based on Markov Chains space [4]. Borda Count methods consider the rankings in their totality, whereas Markov chains allows to model pairwise ranking information. All the items presented in the rankings (or in the union in case of top-k lists) are represented in a graph, where the transition probabilities from one node x_i to another node x_j encode the pairwise ranking information. The computation of the stationary distribution on the Markov chain will determine the aggregate rankings, sorting the node with respect to the probability found at the steady state. The way to assign the initiaz probability matrix is a key point of this approach, in fact four Markov Chain schema (named MC1, MC2, MC3, MC4) has been proposed in [4] and each one uses a different heuristic. Without describing the details of each MC schema, the most interesting for the partial ranking is MC4, where the initial transition matrix gives an high probability value to a move from the state P to the state Q if in the majority of the rankings the items P is ranked above Q. Being heuristic none of the proposed Markov chains methods produces a Kemeney optimal aggregation, but they show interesting performances in practice. [4]. Other Markov chains methods for ranking aggregation has been proposed also in [1], in order to best suit the bioinformatics applications.

Among the heuristic methods, there is also the MEDrank algorithm [17] that is based on the idea of aggregating the input rankings using a median rank for each item. MEDrank in the case of total rankings can optimize the Footrule optimal aggregation if the aggregate ranking has no ties. Moreover, MEDrank satisfies also the Kemeny optimal aggregation within a constant bound (see. [4]). The algorithm is described as follows. Given N total rankings $\tau^1, \dots, \tau^j, \dots, \tau^N$ defined over a set $X = \{x_1, \dots, x_i, \dots, , x_n\}$ of items, let $c(i, \phi)$ a function that returns the number of rankings for which $\tau(i) = \phi$. At the beginning a rank $M(i) = 1$, $i \in X$ is assigned to all the items. The algorithm starts with $\phi = 1$, and at each step it updates for all the items the $M(i)$ as $M(i) = M(i) + c(i, \phi)$. The first item that reaches $M(i) > \beta$ gets a rank value equal to 1 in the output list and it is no more considered. The second item that reaches the same condition gets rank 2 an so on until all the rank values up to ϕ are assigned to the items. The suggested value of threshold β is $\frac{N}{2}$: an item must be counted at least a numebr of times equal to half the number of input rankings before being placed in the aggregate ranking. In the case of top-k lists, MEDrank terminates when the number of items in the aggregate ranking reaches k.

Probabilistic Models for Ranking Aggregation

In this description of the state of the art approaches to ranking aggregation, we wish to include also the solution based on statistical models of rankings such as the Mallow's model and the Luce's model. Mallow's model [11] is a distance-based model which defines the probability of a permutation according to its distance to the location parameter. Mallow's model has also been extended to deal with partial rankings and other distance functions [12][13]. An unsupervised learning algorithm [5] based on the extended Mallows's model has been proposed for the ranking aggregation of total and partial rankings. In this work, an EM algorithm is proposed to learn the model parameters without supervision, and good results has been showed in case of top-k ranking aggregation. Despite the good aggregation results, the computation of the Mallow's model has an high computational complexity ($O(n!)$), which leads to the need for solutions based on it to dominate the complexity as much as possible. Luce's model [9][10] is a stagewise model, where the probability of permutation (of n items) is decomposed in n steps. At each step the model computes the probability for an element to be in any of the n positions through the use of a score assigned to each element. The product of the probabilities retrieved in all the steps, is the probability of the permutation. This model is more efficient in terms of computation time (polynomial) with respect to Mallow's, but it needs a specific score function whereas

Mallow's model is based only on the ranking distance. In order to overcome the limitations and inherit the expressiveness of the two models, a novel probabilistic model called coset-permutation distance-based stagewise model (CPS), has been proposed and evaluated on a ranking aggregation task based on a meta-search problem [19]. The latter is a supervised algorithm that learns the parameters of the CPS model using a training set of rankings, after which an inference step produces the results on a test set.

2.4 Related Works: Learn to Rank Problem

Another issue related to rankings is the learn-to-rank problem. Here the goal is to learn a ranking model from training data. Learn to rank solutions are widely applied to the meta-search problems [3], where the goal is to learn a ranking function that orders the query-documents on the base of their relevance. Two well-known algorithms, based on a pairwise approach are RankBoost [24] and RankSVM [25]. For a complete review of the problem and the algorithms proposed we refer to a specific work [26]. The main difference between the learn-to-rank problem and ranking aggregation is the use of a training set to learn the rank model. This training set includes the rankings but also the relevance values associated with the items. Moreover, in meta-search problem the training data contains also a vector of numeric features relative to the query-document pairs [3]. The ranking aggregation methods presented above do not use relevant labels on the items but only rankings and they do not admit a learning step (expect for the probabilistic models). The ranking aggregation methods presented are total unsupervised solutions in fact no training rankings are available, thus it is not possible to compare the performance of this two approaches.

2.5 Ranking Aggregation vs. The Estimation of True Ranking

As mentioned above, in ranking aggregation the problem is to find a ranking that minimizes the distance from the input rankings. In this work, we deal with ranking aggregation from a point of view that is slightly different from the methods presented in our review of the state of the art. We have noted that the quality of input rankings with respect to a “true ranking’́ and also the existence of the true ranking is not taken into account in the ranking aggregation problem.

The relation between ranking aggregation methods and true ranking has been investigated in [27], where several ranking aggregation methods are compared in order to measure how the quality of the input rankings impacts the performance of the aggregate rankings. The author has showed with rigorous experimental evaluation that the performance of different aggregation methods is deeply connected to the extend whith which the input rankings are related to the true ranking.

Even if a “trueränking is not known in many real situations this does not exclude its existence. This the case of rankings that come from bioinformatic experts, because of the underlying physical reality of the unknown biological phenomenon at hand. An example is the case of microRNA predictor targets [28] where there are a huge number of putative targets and a few number of true targets. The task is to aggregate the output list of the predictors in order to find a consensus list that contains in the first positions the true positive targets. A true ranking exists, but we know only a dichotomous ranking where items should be in top positions if they are true targets, or in the later part of the ranking otherwise [28]. Inspired by these considerations, we tacke the ranking aggregation problem by assuming the presence of a true ranking on given a set of items. The goal is to find a satisfying estime of the unknown true ranking given a set of input rankings provided by experts with different "approximation quality.: The main difference with respect to the ranking aggregation solutions based on minimization criteria is that we assume that the true ranking over the set of items does exist. Since we do not search a consensus list from the input lists, our approach does not address the ranking aggregation of user preferences.

2.6 Belief Function Theory

The Belief Function theory provides a robust framework for reasoning with imprecise and uncertain data, allowing the modeling of subjective knowledge in a non Bayesian way. The theory of the Belief Functions, also known as Dempster-Shafer theory, is based on the pioneering work of Dempster [29] and Shafer [30]. More recent advances of this theory has been introduced in the *Transferable Belief Model* (TBM), proposed by Smets [31]. The Belief Function Theory is a powerful framework to deal with decisions in all the situations where data is imprecise and the subjective views is an important features such as in information fusion tasks. Belief Functions theory generalizes both Set theory (intersection, union) and Probability theory (marginalization, conditioning). The TBM framework is divided in two levels, the *credal level* is where the belief is assigned to a set of possible choices and this belief is updated and combined through several operators, and the *pignistic level* where decisions are taken on the set of choices. In the next sections we provide a basic explanation of the framework through the most common operators and a brief discussion of the application of belief functions on machine learning tasks.

2.6.0-A Representation of Evidence

We define $\Theta = \{\theta_1, \ldots, \theta_k\}$ as a set of propositions about the exclusive and exhaustive possibilities in a certain domain. Θ is called the frame of discernment. Let 2^Θ denote the set of the possible subsets of Θ. A function $m : 2^\Theta \to [0, 1]$ is called *basic belief assignment* (bba) if it satisfies:

$$m(\emptyset) = 0 \qquad \sum_{A \subseteq \Theta} m(A) = 1$$

We call focal elements each subset A of Θ such that $m(A) > 0$. The value $m(A)$ represents the exact belief in the A hypothesis where A can also be a non atomic hypothesis. In this case $m(A)$ represents the belief that supports the set A and it makes no additional claims to any subsets included in A. We introduce also the normal condition of bba related to the belief assigned to $\emptyset$. A normal bba has $m(\emptyset) = 0$ whereas $m(\emptyset)$ is focal set for sub-normal bba. TBM permits also an open-world assumption and manage also sub normal bba (normalization operation), in this work we deal only with normalized bba. The belief assigned on $m(\emptyset)$ can assume two different interpretations, one is the conflict after the combination of several experts, the other is that the expert (or the combination of the experts) has belief in an hypothesis outside the frame Θ. The modeling of the ignorance of an expert lies in the possibility to assign evidence to a set of elements instead of assigning it to just a single element. If $\Theta = \{a, b\}$ and in the case the expert does not have any prior knowledge, $m(\{a, b\}) = 1$ represents the total ignorance or confusion of the expert to decide between the two events. In the probability framework the uncertainty could be modeled using a prior over the events. In the previous example the ignorance/confusion could be modeled as $P(a) = P(b) = 0.5$ where the two propositions a and b have the same prior probabilities. It is easy to recognize the differences of the two approaches respect to the modeling of the uncertainty. In the probability model we have probability values for the two propositions whereas in the Belief Function framework the model directly represents the total ignorance over the pair of propositions without any additional assumption. Some of the possible bbas are:

Bayesian All focal elements are singletons ($\Theta = \{a, b\}$, $m(a), m(b) > 0$)

Simple The bba has two focal sets and one of those is Θ ($\Theta = \{a, b\}$, $m(a) > 0$ and $m(\Theta) > 0$)

Categorical The bba has only one focal set ($\Theta = \{a, b\}$, $m(a) > 0$)

Vacuos The bba has only $m(\Theta) = 1$ as focal set.

If m is a valid bba ove the frame Θ, then the belief function $Bel : 2^{\Theta} \rightarrow [0, 1]$ is defined as:

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

Another notion introduced in this framework is the plausibility function. $Pl : 2^{\Theta} \rightarrow [0, 1]$ defined as:

$$Pl(A) = \sum_{A \cap B \neq \emptyset} m(B)$$

It is also possible to express the plausibility as $Pl(A) = \sum\limits_{A \cap B \neq \emptyset} m(B)$. The quantity $Bel(A)$ is the degree to which the evidence supports A, whereas the $Pl(A)$ is the upper bound of the degree of support that could be assigned on A. Moreover, it is possible to obtain m from the Bel via the following trasformation:

$$m(A) = \begin{cases} \sum\limits_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel(B) & A \neq \emptyset \\ 1 - Bel(\Theta), & A = \emptyset \end{cases} \tag{2.8}$$

we point out that m, Bel, Pl are equivalent representations of a piece of evidence. In case of Bayesian bbas, it can be shown that Pl=Bel, that Belief functions and Bayesian approaches are equivalent.

2.6.0-B Combination and Updating the Belief

In order to aggregate distinct sources $m_1, \dots, m_n$ on Θ, the framework provides several combination rules, such as the conjunctive rule, the disjunctive rule and the caution rules [31] among others. The conjunctive rule is defined as:

$$m_{1\textcircled{\cap}2}(A) = \sum_{B \cap C = A} m_1(B) m_2(C) \quad A \subseteq \Theta \tag{2.9}$$

The conjunctive rule $\textcircled{\cap}$ is justified when all the sources of belief are supposed to assert the truth and to be independent. Moreover, $m_{1\textcircled{\cap}2}(\emptyset)$ represents the degree of conflict between the two bbas. Conflict arises when the different sources have singleton focal elements, in which case their intersection is $\emptyset$. Another conjunctive operator is *Dempster's combination rule*, defined as:

$$m_{1\otimes 2}(A) = \begin{cases} \frac{\sum\limits_{B \cap C = A} m_1(B) m_2(C)}{1-K} & A \neq \emptyset \\ 0 & A = \emptyset \end{cases}$$

where K is the degree of conflict $m_{1\textcircled{\cap}2}(\emptyset)$: the $\otimes$ operator the conflict is used to normalize the combined bbas. The conflict in the belief function framework can have different meanings and it should be managed in accordance to the application at hand. One possible meaning is that the frame Θ is not exhaustive, and $m_{1\textcircled{\cap}2}(\emptyset)$ quantifies the belief that there exist hypothesis θ outside the frame Θ (open-word assumption). Otherwise $m_{1\textcircled{\cap}2}(\emptyset)$ means that the sources do not report on the same object, this is applied in some applications where the sources are clustered according to which object they report [32]. The use of the two conjunctive operators $\otimes$ and $\textcircled{\cup}$ is related to the type of application. The conjunctive operator is well indicated when we need to keep track of the conflict between the bbas to combine, wheres the $\otimes$ rule is suggested when the conflict should be normalized.

If the sources to combine are still independent but at least one of the tells the truth (without knowing with one), then the disjunctive combination rule is more appropriate. Given two mass functions m_1 and m_2 defined on Θ the disjunctive rule $\textcircled{\cup}$ is defined as:

$$m_{1\textcircled{\cup}2}(A) = \sum_{B \cup C = A} m_1(B) m_2(C) \quad A \subseteq \Theta \tag{2.10}$$

The disjunctive operator is associative, commutative and admits as neutral element the bba which assigns a total belief to the empty set ($m(\emptyset) = 1$). In the literature other combination rules have been proposed such as the Debois and Prade's rule [33] and the Yager's rule [34]. These combination rules are a mix of the conjunctive and disjunctive operators, and propose other ways to deal with the conflict generated by the combination. Even if we do not present the details, we mention also the cautious rule [35], a combination rule appropriate when the sources of belief are dependent.

The discount operation faces the situation when the information used by experts are not fully relevant or reliable due to the presence of faulty sensors or other a priori information depending on the problem. Given $\alpha \in [0, 1]$ the degree of reliability of the expert, the discount operation is defined as [36]:

$$\begin{aligned} m^*(A) &= \alpha m(A) \quad A \subset \Theta \\ m^*(\Theta) &= 1 - \alpha[1 - m(\Theta)] \quad A = \Theta \end{aligned} \tag{2.11}$$

where m^* are the discounted bbas. When decreasing α down to 0 the bba loses all its information, and the result is a vacuous belief function $m(\Theta) = 1$. A fully realiable source has $\alpha = 1$, which leaves the bba unchanged. In case of simple bbas over a frame Θ defined as follows,

$$\begin{aligned} m(A) &= s \\ m(\Theta) &= 1 - s \\ m(B) &= 0 \quad \forall \quad 2^{\Theta}\{\Theta, A\} \end{aligned}$$

where $m(A)$ is the only focal element except $m(\Theta)$, and $s \in [0, 1]$, the discount operation can be rewritten as:

$$\begin{aligned} m^*(A) &= \alpha s \\ m^*(\Theta) &= 1 - \alpha s \end{aligned} \tag{2.12}$$

2.6.0-C Decision Making

In the TBM framework the uncertainty reasoning is performed at the so-called credal level, where the bbas are combined and updates, instead the decision making is made at the pignistic level where the belief are used to make decision [31][37]. TBM framework to make decision requires to quantify the belief in probability in order to avoid the Dutch Books. Dutch Book is a set of bets that lead to a sure loss regardless of the outcome of the gamble. This only way to prevent this situation is to be certain that our belief is reppresented by a probability function. Smets proposes the *pignistic transformation* to transforms the masses defined on the power set of Θ to a probability space defined only on single atoms of Θ as following:

$$Betp(\theta) = \sum_{\{A \subseteq \Theta, \theta \in A\}} \frac{m(A)}{1 - m(\emptyset)|A|} \quad \forall \theta \in \Theta \tag{2.13}$$

The idea underling the pignistic transformation is to distribute equally every bba on the singleton elements that belongs to its focal element. Smets justifies the pignistic trasformation as the only trasformation that satisfies five specific assumptions related to the properties that the trasformation must satisfy. For a detailed description of all the assumptions satisfied by the pignistic trasformation we refers to [38]. Even if in this work we use the pignistic trasformation, other methods to map bbas into a probability measure such as the plausibility trasformation [39] has been proposed in literature.

In this introduction we have introduced the basic concepts of the frameworks that are related with the application of the Belief Function in our solution. All the concepts covered are found in the various works on the Belief functions [31][40] and Dempster-Shafer Theory [29][30].

2.7 Applications of Belief Function Theory

The Belief functions has been in applied to several different applications such as audit risk [41], decision making problems [42] and information fusion [36]. To the best of our knowledge, the belief function framework has not yet applied to the ranking aggregation problem even if the ranking aggregation problem shares same aspects of the information fusion problems. We focus our discussion on the belief functions regard to the machine learning tasks because it is the more related field with the ranking aggregation problem. With respect to machine learning problems, the Belief Function Theory has been applied to different tasks such as classification, clustering and the combination of classifier output. In [43] Denoaux *et al.* propose the evidential k-nn rule, a nearest-neighbor algorithm based on the belief function. Each neighborhood of the instance to be classified supports a piece of evidence relative to the class membership. This support is quantified by the euclidean distance between the two points. Finally the evidence of k nearest neighborhood are combined through the Dempester's rule. An extension of the evidential k-nn rule [44] has been applied to the task of multi-label classification, where the examples can belong to several classes and not only a single one. With regard to the clustering problem, several solutions based on belief function have been proposed [45][46][47]. One of the possible drawbacks that limits a large spread of this framework, is the high computational cost due to the fact that all the operations (combinations) works in the power set of the frame Θ. On the other hand, this complexity gives to the belief function a powerful tool to model detailed situations of imprecise data. A solution provided to overcome the high computation cost, is to constraint the focal elements in intervals that can be represented in any lattice structure not necessary with linear order [48].

Chapter 3

BRE: Belief Ranking Estimator

3.1 Introduction

The ranking aggregation problem arises when it is necessary to combine different rankings on a finite set of items, in order to produce a new ranking that satisfies specific criteria. Usually, this corresponds to the necessity to combine the opinion of experts with different background, such as the combination of ranked lists of differently expressed genes provided by different microarray analysis methods, or the combination of search engine results [4][3], or committee decision making. Most of the methods proposed for the combination of rankings aim to minimize the distance among the input rankings for a given ranking distance. This is the case of the Footrule optimal aggregation [4], and the cross-entropy method [14] to approximate the Kendall optimal aggregator. Other aggregator methods are based on heuristics such as the Borda Count [16] methods (that includes the mean and median as aggregation functions), MEDrank [17] and Markov Chain methods [4]. Despite the presence of a true ranking is overlooked in the formulation of the ranking aggregation problem, the relation between ranking aggregation methods and the true ranking has been investigated in a work [27] that shows how the results of the aggregation methods are affected by the noise of the input rankings with respect to the true ranking. In this work ranking aggregation is tackled assuming the existence of a true ranking of the underlying set of items. The goal is to find a satisfying estimation of the unknown true ranking given a set of input rankings provided by experts with different approximation quality. The main difference with respect to ranking aggregation solutions based on minimization criteria is that we assume that the true ranking over the set of items does exist. We claim that this is the case when the rankings comes from bioinformatic rankers because of the underlying physical reality of the unknown biological phenomenon at hand. Our solution to the ranking aggregation problem is based on the Belief Function Theory that provides a solid framework for reasoning with imprecise and uncertain data, allowing for the modeling of subjective knowledge in a non Bayesian way. The application of Belief Function to rankings gives the possibility to encode different *a priori* knowledge about the correctness of the ranking positions and also to weight the reliability of the experts involved in the combination. Moreover, to the best of our knowledge the use of Belief Function on ranking aggregation problem has not been proposed yet in literature. Our algorithm, called Belief Ranking Estimator (BRE), estimates the true ranking in an unsupervised way given a set of input rankings. We evaluate BRE on total rankings of synthetic data and compare our method against some ranking aggregation competitor methods.

3.1.1 Notation and Definition of the Problem

Let $X = \{x_1, \ldots, x_n\}$ be a set of items to be ranked by an expert opinion. We denote as $\tau = (\tau(1), \ldots, \tau(n))$ a ranking associated to X, where $\tau(i)$ is the rank associated to the item x_i. We suppose to have $\tau^{T_{rank}} = (\tau^{T_{rank}}(1), \ldots, \tau^{T_{rank}}(n))$, that is the golden "true‘ ranking on the items of X, and we denote as R_j the expert involved in the ranking, so for each expert we have a corresponding ranking $\tau^{R_j} = (\tau^{R_j}(1), \ldots, \tau^{R_j}(n))$. To simplify the notation, each ranking is denoted by τ^j for all $j \in N$ where N is the number of experts and consequently of rankings. We suppose also that the most important items for a ranking τ^j receives a rank value equal to n. This assumption in the set of the permutations does not lead to any loss of generality. The problem in its general form is stated as follows. Given N rankings τ^{R_j} of length n of the n items $X = \{x_1, \ldots, x_n\}$, namely permutations, that estimate with unknown quality the unknown true ranking $\tau^{T_{rank}}$ find a ranking that estimates the true ranking.

3.2 BRE: Belief Ranking Estimator

The Belief Ranking Estimator (BRE) is an unsupervised algorithm that iteratively computes an estimation of an unknown true ranking, given N input rankings that are assumed to be approximations of unknown quality of the true ranking. The core of the method is the use of belief functions in order to capture and model the uncertainty regard the position of each item contained in each ranking. Through the use of a true-rank estimator, BRE estimates the quality of the input rankings to use this information in the combination process. The main steps of the Belief Ranking Estimator are the following:

- Mapping the rank value of each item into belief assignments.
- Assessment of the quality of the input rankings using a true-rank estimator.
- Application of the quality information of the input rankings to the belief assignments .
- Combination of the beliefs associated with each item to produce a ranking as outcome.
- Iteratively replacement of the worst ranking with the combined ranking produced.

We present two versions of BRE, the iterative version that includes the quality information of the input rankings as weight in the combinaton step (Alg. 1) and the not weighted version that combines the belief distribution of the input rankings without the application of the weights (Alg. 2).

Algorithm 1 Belief Ranking Estimator: Iterative version

```
input I=$\tau^1, \dots, \tau^N$ // a vector of N Rankings
input $T$ // Numbers of iterations
input $TE$ // True-rank estimator
  k= 0
  BE=Belief_From_Rankings(I)
  $FinalRank_k$=Combination(BE)
  while k != $T$ do
    $\bar{w}$=ComputeWeights(I,TE(I))
    BE=ApplyWeights($\bar{w}$,BE)
    $FinalRank_k$=Combination(BE)
    I[pos(max($\bar{w}$))]=$FinalRank_k$
    BE=Belief_From_Rankings(I)
    k++
  end while
output $FinalRank_k$
```

3.2.1 BBA From Rankings

The mapping from rankings to belief assignments expresses all our *a priori* knowledge about the experts involved in the combination. Since we do not have at hand a specific application context with wide *a priori* information, we assume to use only the rank values associated to each element. Notice that in a ranking the highly-considered items may have high or low rank values. Both cases are correct but this information should be considered to produce the right mapping according to the interpretation of the input rankings.
We consider a simple frame of discernment $\Theta = \{P, \neg P\}$, where P, $\neg P$ are the hypothesis that an element is ranked in the right position or not respectively. The bba definition should reflect the fact that high-ranking elements have more belief to be in the right position from the point of view of the expert who provided the ranking. Since the lack of external information about the correctness of the ranking we are not able to assert if an element is not in the right position ($\neg P$), the remaining belief is assigned to the uncertainty between the two possible hypotheses, namely to Θ. Given a set of N rankings $\tau^1, \dots \tau^j, \dots, \tau^N$ of the same n elements, the bba of the j-th ranking on the i-th element is consequently assigned as:

$$\begin{aligned} m_{ji}(P) &= \frac{\tau^j(i)}{n} \\ m_{ji}(\neg P) &= 0 \\ m_{ji}(\Theta) &= 1 - \frac{\tau^j(i)}{n} \end{aligned} \tag{3.1}$$

We use the above assignment in case of rankings where high-rank values correspond to the highest positions. In the case of rankings where low-rank values correspond to the

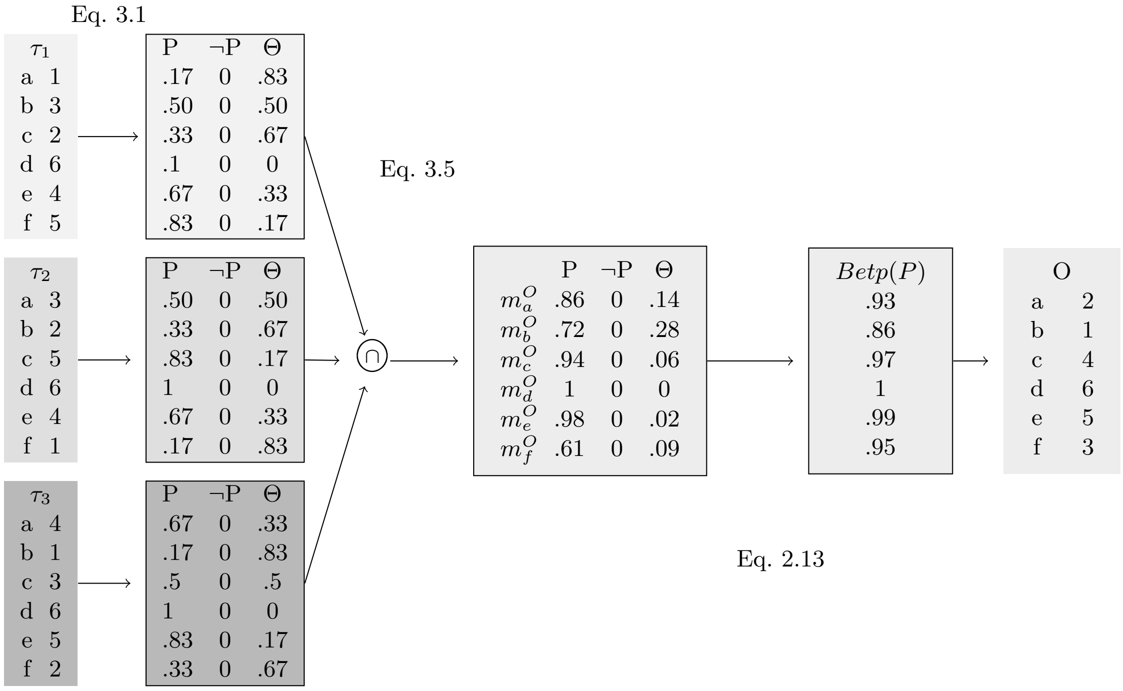


Figure 3.1: Example of BRE with NW schema: BBA from rankings (Eq. 3.1), combination (Eq. 3.5) and the ranking outcome (Eq. 2.13)

highest position, the alternative but equivalent belief assignment is:

$$\begin{aligned} m_{ji}(P) &= \frac{n-(\tau^{j}(i)-1)}{n} \\ m_{ji}(\neg P) &= 0 \\ m_{ji}(\Theta) &= 1-\frac{n-(\tau^{j}(i)-1)}{n} \end{aligned} \tag{3.2}$$

where 1 is the lowest values present in the rank. We have used Eq. 3.1 in our experiment, however we have reported both to highlight the equivalence of the two interpretations of the ranking in terms of bba on Θ. More complex assignments will be discussed and evaluated in the case of partial rankings. The bba proposed above are computed in the *Belief_From_Ranking* routine in both NW version (Alg. 2) and in the iterative version (Alg. 1). An numerical example of the bba proposed in Eq.3.1 is showed in Fig. 3.1.

3.2.2 Weight Computation

As quality of the input ranking, we mean how the input rankings are informative with respec to the true rankings Since the true ranking $\tau^{T_{rank}}$ is not available to estimate of the qualities of the input rankings by the unsupervised context of the problem, we introduce an estimator (TE) of the true ranking as input in order to assess the quality of the rankings. Let denote τ^{TE} a ranking produced by the estimator TE from the input rankings. The weight of the j-th ranking is computed with the Spearman footrule distance

Algorithm 2 Belief Ranking Estimator: Not weighted version

```
input I=$\tau^1,\dots,\tau^N$ // a vector of N Rankings
input $T$ // Numbers of iterations
  BE=Belief_From_Rankings(I)
  $FinalRank_k$=Combination(BE)
output $FinalRank_k$
```

normalized as:

$$w_j = \frac{F(\tau^j, \tau^{TE})}{\frac{1}{2}n^2} \quad \forall j \in 1,\dots,N \tag{3.3}$$

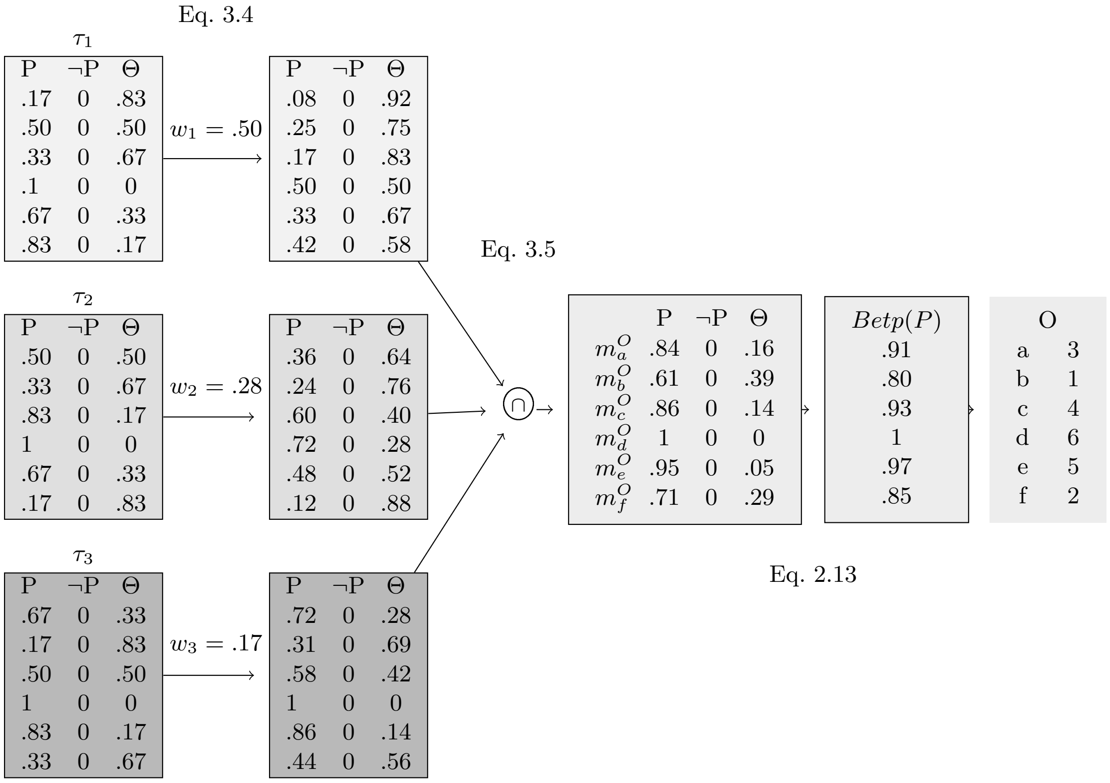
where $F(\cdot\,,\cdot\,)$ is the Spearman footrule distance [4] (Sec. 2.2) defined over two rankings τ, σ as:

$$F(\pi,\sigma) = \sum_{i=1}^{n} \mid \pi(i) - \sigma(i) \mid$$

In order to obtain weight values in the interval $[0,1]$, the distance F is divided by the maximum values of the Spearman footrule distance for rankings of length n [6]. For two identical rankings w will be 0, instead of $w = 1$ that corresponds of two totally-inverted rankings. By $\bar{w}$ is denoted the vector of the weights computed for all the N rankings.
As an estimator it is possible to use any ranking, even a fixed raking based on some *a priori* knowledge of the problem. Given the unsupervised nature of BRE, we derive the estimator ranking by the aggregation of the input rankings through the methods presented in Chapter 2. The more the estimator ranking is a good approximation of the underlyng true ranking, the more the weights will be effective to represent the actual quality of the input rankings. Other distances among rankings, such as Kendall [4] and Coset-permutation distance [19] are still valid to compute ranking weights inside our method. In this work we have tested only the Spearman footrule distance, since we have focused our work to study the role of the Belief Function theory on this unexplored application context. The weights computation is executed by the *ComputeWeights* routine in Alg. 1.

3.2.3 Application of the Weights

In the Belief Function framework, the discount operation aims to reduced the belief assignment on the frame with respect to the degree of reliability of the source as showed in Eq. 2.11. In BRE we propose an operation slightly different from the original disocunt, in the sense that it increases also the belief for the most important sources. Our idea of bba discount is to reduce the uncertainty between P and $\neg P$, proportionally to the correspondent weight for the best ranking, and to increase the uncertainty for all the other rankings. Even if the operation defined is not the same of the original discount operation, we refer to it as the application of the weights. The discount of the bba of each ranked



τ_1

P	¬P	Θ
.17	0	.83
.50	0	.50
.33	0	.67
.1	0	0
.67	0	.33
.83	0	.17

P	¬P	Θ
.08	0	.92
.25	0	.75
.17	0	.83
.50	0	.50
.33	0	.67
.42	0	.58

τ_2

P	¬P	Θ
.50	0	.50
.33	0	.67
.83	0	.17
1	0	0
.67	0	.33
.17	0	.83

P	¬P	Θ
.36	0	.64
.24	0	.76
.60	0	.40
.72	0	.28
.48	0	.52
.12	0	.88

τ_3

P	¬P	Θ
.67	0	.33
.17	0	.83
.50	0	.50
1	0	0
.83	0	.17
.33	0	.67

P	¬P	Θ
.72	0	.28
.31	0	.69
.58	0	.42
1	0	0
.86	0	.14
.44	0	.56

	P	¬P	Θ
m_a^O	.84	0	.16
m_b^O	.61	0	.39
m_c^O	.86	0	.14
m_d^O	1	0	0
m_e^O	.95	0	.05
m_f^O	.71	0	.29

Betp(*P*)
.91
.80
.93
1
.97
.85

O	
a	3
b	1
c	4
d	6
e	5
f	2

Figure 3.2: Example of BRE with weighting schema: weight application (Eq. 3.4), combination (Eq. 3.5) and the ranking outcome (Eq. 2.13)). The weights are computed using the mean of the rankings as true-rank estimator (Eq. 3.3).

element is described as follow:

$$\begin{array}{ll}
if \quad w_j = min(\{w_1, \ldots, w_N\}) & if \quad w_j \neq min(\{w_1, \ldots, w_N\}) \\
m'_{ji}(P) = m_{ji}(P) + (w_j * m_{ji}(\Theta)) & m'_{ji}(\Theta) = m_{ji}(\Theta) + (w_j * m_{ji}(P)) \\
m'_{ji}(\neg P) = 0 & m'_{ji}(\neg P) = 0 \\
m'_{ji}(\Theta) = 1 - m'_{ji}(P) & m'_{ji}(P) = 1 - m'_{ji}(\Theta)
\end{array} \tag{3.4}$$

where m_{ji} is the bba of the j-th ranking on the i-th item, $min(\cdot)$ is the minimum function and m'_{ji} is the discounted one. We apply these weights globally, namely the weight value is applied to all the bba's items of each ranking. The discount operation is showed in Fig. 3.2 where the mean of the rankings is used as estimator. The idea is to reduce the uncertainty, proportionally to the correspondent weight for the best rankings (namely, the ranking with minimum weight), and to increase the uncertainty for all the other rankings. Note that the bba of $\neg P$ are not modified, since new evidence regard to the items ranked in wrong positions is not added. The application of the weights is consistent within the framework, since the sum of bbas in the frame Θ is still 1 for each item. This operation is done in the *ApplyWeights* routine in Alg. 1. A numerical example of the application of the weights is showed in Fig. 3.2.

3.2.4 Ranking Output

The final step of BRE is the combination of the bba of each item along all the rankings, using the conjunctive rule (Eq. 2.9) as follows:

$$\begin{aligned} m_i^O(P) &= \textcircled{\cap}_{j=1}^{N} m_{ji}(P) \\ m_i^O(\neg P) &= \textcircled{\cap}_{j=1}^{N} m_{ji}(\neg P) \\ m_i^O(\Theta) &= \textcircled{\cap}_{j=1}^{N} m_{ji}(\Theta) \end{aligned} \tag{3.5}$$

with $i \in 1, \ldots, n$ and where m_i^O is the combined bba for the i-th item. The use of the conjunctive rule is justified when all the sources of belief are assumed to tell the truth and to be independent. These requirements are fully satisfied here, since we suppose that the rankings are independent and totally reliable because the unsupervised context does not allow to make other assumptions on their quality. We apply the Eq. 2.13 on the m_i^O in order to take decisions in the frame Θ. The final ranking $O = (O(1), \ldots, O(i), \ldots, O(n))$ is produced by sorting all the items with respect to $BetP_i(P)$, that corresponds to the probability of the i-th item of being in the right position. The combination step is done in the *Combination* routine in both NW version (Alg. 2) and the iterative version (Alg. 1).

3.2.5 BRE Versions

As described before, the three parts described above are embedded inside an iterative procedure that aims to replace the worst rankiing with combined ranking produced during each step as showed in the Alg. 1. The idea underlying this iterative replacement is that the ranking computed in each iteration will be more informative instead of the input rankings in terms of approximation quality of the true ranking. For the replacing of a possible good true-rank estimator, we expect that BRE will increase the quality of the true-rank estimator. The effect of the iterative procedure in the algorithm will be evaluated in details in the experimental parts (Sec. 3.4-3.5). Although there is no theoretical constraint about the number of iterations, we propose as number of iteration $MAXT = \frac{N}{2}$. The rational of this rule of the thumb is that replacing more then one half of the original rankings can possibly lead to poor performance due to the replacement of some of the best rankings with information affected by the worse ones.

We present three versions of BRE, one not-weighted (BRE-NW) where the rankings quality is not involved in the combination, the iterative version where weights and the ranking replacement are introduced and a $T = 1$ version without replacement. $BRE - NW$, showed in Alg. 2, combines the belief distribution of the input rankings without the application of the weights. A numerical example of the BRE-NW is showed in Fig. 3.1. In the remaining of the chapter we refer as weighting schema to the $BRE-1T$ version, whereas as iterative schema when $T = MAXT$.

In our solution we assume to use only the rank values associated to each elements, but BRE could be easily adapted to the case where other information is available to the specific ranking aggregation problems. This *a priori* knowledge can include the truthfulness of experts and the reliability of some rank items with respect to other items. The truthfulness of the experts concerns the global reliability of the rankings, it is also related to the quality of the rankings with respect to the true ranking. If this information is available *a priori* can be directly used in BRE as input weights.

Another issue related to the *a priori* information available, is the possibility to known the information about each ranked item or for a subset of items. An example can be an expert that have a bias on the ranking of a subset of elements so it assigns systematically higher or lower rank values to these items in the ranking produced. Starting from this knowledge, other belief assignments should be considered, in order to map this bias into the frame Θ. A more complete scenario is when the information of the items are known *a priori* for all the rankings, this can be perfectly managed into the BRE algorithm supported by the Belief Function that permits to model the subjective point of view for the item of each ranking involved. The last consideration opens also the possibility to use and compute the weights for each item instead of a global weights applied to all the items of a ranking.

Although we proposed and explored the BRE algorithm on a simple scenario, the above issues has been mentioned to highlight how BRE can handle the different *a priori* knowledge by to the use of the Belief Function.

3.3 BRE vs. the Ranking Aggregation Methods

With respect to the classification in heuristic methods and optimization solutions for the rankings aggregation problem we can consider BRE to be an heuristic solution, since in its formulation there is no criterion to minimize. Among the state-of-the-art methods presented in Chapter 2, in the next experiments we have compared the performance of BRE with respect to the following methods: the mean and the median of the rankings (Borda Count's method) and the Footrule optimal aggregator. We point out that BRE uses the Spearman footrule distance for the computation of the weights, so to provide a fair comparison we focus on solutions that minimize that distance. As heuristic competitors we use the Borda Count methods with the median and the mean as aggregation functions, where the score of each item corresponds to its rank value ($B^J(i) = \tau^j(i)$). We simply refer to Borda Count's methods as the mean and the median of the rankings. As for the optimal aggregator method, we include as competitor the Footrule optimal aggregator. The Footrule optimal aggregator minimizes the footrule distance with the input rankings, and it can be computed in polynomial time solving the minimum cost of matching on a weighted bipartite graph [4].

We have not included the MEDrank algorithm as competitor since we have just evaluated similar heuristics as the median of the rankings. Moreover the MEDrank algorithm provides the Footrule optimal aggregator in case of total rankings [17], and we have just included a similar solution as competitor.

The Markov chain methods take into account in their solutions the pairwise comparison of all the items on the rankings. We notice that the Markov chain solutions consider the problem from the point of view of the items present in the rankings. On the other hand, BRE faces the problem from the the point of the input rankings in fact the informations related to the pairwise comparison of all the items are not used. For the highlighted differences of the two approaches, we have not included the Markov chain methods as competitors.

The stochastic optimization solutions are not included as competitors on total rankings, since the Footrule optimal aggregator on total rankings is solved with acceptable computational time. In general, BRE is not compared with the rankings aggregation methods based on probabilistic models since they include a step (both unsupervised and supervised solutions) where the probabilistic model learns from data. Moreover BRE and the other ranking aggregation methods are totally unsupervised solutions.

3.4 Experiment 1: BRE vs The Competitor Methods

In this section we describe the results of BRE with respect to some aggregation methods proposed in the state of the art. All the versions of BRE previously presented such as BRE-NW, BRE-1T and BRE-MAXT have been evaluated in order to highlight possible differences of performance.

The goals of this experiment is to evaluate the performance of BRE throught different cases of quality of input rankings and the evaluation of BRE with respect to the mean, the median and the Footrule optimal aggregator (Opt_list) [4]. A detailed description of the

aggregation methods evaluated has been given in Sec. 2.3. As true-rank estimator inside BRE we use the same ranking-aggregation competitor methods, in order to investigate if BRE increases the performance with respect to the methods used as true-rank estimator. We have not found real data on total rankings with an available true ranking. To the best of our knownledge, there is not any. For this reason we have decided to evaluate BRE on synthetic data that suits perfectly the problem at hand.
The data has been generated as follows. We have fixed a true ranking ($\tau^{T_{rank}}$) from which the input rankings has been randomly generated according to fixed values of the Spearman coefficient, indicated as ρ [20][22] (Eq. 2.4). The generated rankings are overall random permutations with respect to all the items contained in the true ranking. The variables that would be investigated in our experimental settings on synthetic data are the following:

– Correlation ρ of the input rankings with respect to true ranking.

– Number of experts (denoted as N).

Despite the space of the parameters is quite huge to be totally evaluated, we have decide to fix the length of the rankings ($n = 300$) in order to focus our attention on the correlation of the rankings (a measure of quality) and on the number of experts aggregated. Among all the N values we have generated a total of 10 different cases that permit to have a large picture of the BRE performance in heterogeneous situations of correlation with respect to the true ranking. The length of the ranking n has been fixed equal to 300 for all the cases. We have evaluated a number of experts N equal to 3, 10, 30. For each N value different cases of the input rankings has been proposed. For N=3 we have defined 4 cases:

Case 1 1 ranker extremely good (ρ=.80) with respect to the others (ρ=.06 ρ=.01).

Case 2 two good rankers ($\rho = .60, .40$) and a very poor one (ρ=.01).

Case 3 3 rankers with high correlation (ρ=.80, .60, .10).

Case 4 3 rankers with poor correlation (ρ=.01, .06, .03).

For N = 10, 30 we have defined 3 cases each:

Case *Good* the 80% of the rankers are highly informative ($\rho \in [.95, .70]$) and the remaining 20% are low correlated ($\rho \in [.30, .10]$).

Case *Equal* The rankers are equally distributed among the three types: highly, medium ($\rho \in [.70, .30]$) and low correlated.

Case *Poor* The opposite of the case *good*, 80% of the rankers are poorly informative and only the 20% are hightly correlated.

For the above cases the ρ values are randomly chosen within the defined intervals. In order to have more reliable results we performed 10 independent replicas of the procedure using the same generation parameters for each case and N value. The statistical significance of the difference of the averages between BRE and its estimators used as competitors is computed with a paired two-tailed t-Test on the 10 replicas (with α = .05). The

Table 3.1: Spearman correlation coefficent (ρ) and Spearman Footrule distance (F) of BRE and of the competitor methods with respect to the true ranking. ● means that BRE is significantly better than the corresponding competitor, and ○ means that BRE is significantly worse.

		Evaluation measure ρ									
Method	True-Rank.	3 Rankers Cases				10 Rankers Cases			30 Rankers Cases		
	Est.	MAXT=3T				MAXT=5T			MAXT=15T		
		1	2	3	4	good	equal	poor	good	equal	poor
Random		.1895	.2664	.5452	.0424	.5341	.5737	.5302	.8028	.4995	.3581
Mean		.4781	.5419	.7958	.0782	.9621	.8760	.7793	.9856	.9543	.8802
Median		.4257	.5106	.7678	.0693	.9748	.8656	.7641	.9941	.9546	.8579
Opt_list		.4065	.4953	.7515	.0594	.9754	.8686	.7681	.9957	.9663	.8787
BRE-NW		.4888	.5254	.7799	.0804	.9383	.8453	.7723	.9409	.8941	.8074
BRE-1T	Mean	.3826	.5742	.8226●	.0722	.9763●	.9207●	.8893●	.9903●	.9742●	.9353●
BRE-MAXT	Mean	.3464	.5780	.8311●	.0666	.9782●	.9270●	.8880●	.9785●	.9714●	.9372●
BRE-1T	Median	.4305	.5865●	.8229●	.0699	.9751	.9208 ●	.8890 ●	.9904○	.9743●	.9342●
BRE-MAXT	Median	.3981	.5915●	.8319●	.0660	.9781	.9276●	.8914●	.9784○	.9709●	.9371●
BRE-1T	OptList	.4717	.5826●	.8234 ●	.0729	.9767	.9212●	.8856 ●	.9904○	.9755●	.9374●
BRE-MAXT	OptList	.4415	.5844●	.8328 ●	.0692	.9783	.9276●	.8919●	.9780○	.9716●	.9391●
		Evaluation measure: F									
Random		.4790	.5270	.3790	.6410	.2550	.4030	.5090	.2650	.4090	.5530
Mean		.4117	.4045	.2828	.5318	.1763	.2858	.3461	.1637	.2639	.3438
Median		.4160	.4057	.2575	.5687	.0673	.2186	.2927	.02290	.1359	.2813
Opt_list		.4551	.4399	.2808	.6311	.0535	.1780	.2523	.0144	.0671	.1660
BRE-NW		.4511	.4393	.2912	.6235	.1444	.2368	.3129	.1444	.1919	.2669
BRE-1T	Mean	.4938	.4132	.2578●	.6262○	.0926●	.1688	.1982●	.0592●	.0936●	.1482●
BRE-MAXT	Mean	.5079	.4108	.2475●	.6282○	.0888●	.1625●	.2014●	.0852●	.1044●	.1525●
BRE-1T	Median	.4704	.4071	.2581	.6275○	.0953○	.1687●	.1983●	.0591○	.0936●	.1493●
BRE-MAXT	Median	.4815	.4044	.2473	.6284○	.0890○	.1619●	.1977●	.0851○	.1057●	.1531●
BRE-1T	OptList	.4488	.4113●	.2576●	.6254	.0919○	.1683●	.2011●	.0590○	.0914○	.1456●
BRE-MAXT	OptList	.4569	.4098●	.2469●	.6264	.0888○	.1618●	.1967●	.0861○	.1043○	.1510●

performance is measured with the Spearman correlation coefficient (ρ, Eq. 2.4) and the Spearman Footrule distance (F) computed with respect to the true ranking ($\tau^{T_{rank}}$).

In Tab. 3.1, we show the significance of the results of BRE with respect to the true-ranking estimator method used. In the discussion the significance has been also evaluated such as BRE-1T vs. BRE-MAXT and BRE-1T vs. BRE-NW.

In Tab. 3.1, we also present a competitor called *random*, that corresponds to a ranking chosen uniformely randomly among the input rankings. Since BRE-1T is significantly better than the random competitor in all the cases and for both the evaluation measures, we can assert that the BRE results are very far from a random guess.

The comparison between BRE with the mean as true-rank estimator and mean as aggregation method, shows that BRE-1T and BRE-MAXT outperform the mean in most of the cases for both evaluation measures (ρ and F), except for the cases 1 and 4 ($N = 3$) where the mean shows higher results. BRE-1T shows significant performance in the majority of the evaluated cases. Regarding the case *poor* for $N = 10$ and $N = 30$, BRE-1T

Table 3.2: Average Spearman correlation coefficent (ρ) and Spearman Footrule distance (F) of BRE and of the other competitors with respect to true ranking for the *good, equal and poor* cases with $N = 10$, 30

		Evaluation measure ρ	
Methods	T.E	10 Rankers	30 Rankers
Mean		.8725	.9400
Median		.8649	.9355
Opt_list		.8707	.9469
BRE-1T	Mean	.9288	.9666
BRE-1T	Median	.9283	.9663
BRE-1T	Opt_list	.9326	.9629
		Evaluation measure F	
Mean		.2694	.2572
Median		.1929	.1467
Opt_list		.1613	.0825
BRE-1T	Mean	.1532	.1003
BRE-1T	Median	.1541	.1007
BRE-1T	Opt_list	.1537	.0987

highlights a significant improvement with respect to the mean in terms of F and ρ, whereas the mean is influenced by the low-quality rankings.

Taking into account the median as the true-rank estimator, we notice that BRE-1T outperforms the median as competitor method in most of the evaluated cases for both evaluation measures, except for the case *good* with N =30 where the median outperforms our BRE. Also for the median, BRE-1T performance shows a significant improvement in the cases *poor* for $N = 10$, 30 with both the evaluation measures.

Opt_list shows the best values of ρ and F with respect to the mean and the median in all the three cases with $N = 30$ and $N = 10$. BRE-1T with Opt_list as estimator outperforms significantly the Opt_list method in all cases *poor* ($N = 10$, 30) , except for the case *good* ($N = 30$, 10) and the case *equal* with $N = 10$ where Opt_List shows the best results among all the other ranking aggregation competitors with F distance.
From Tab. 3.1, we notice that BRE-MAXT shows significant improvement with respect to the estimators in same cases against the 1T version. From the comparison of BRE-1T with respect to BRE-MAXT, we point out that in the cases *good* and *equal* ($N = 10$), BRE-MAXT outperforms significantly the 1T version even if for small differences of ρ and F. On the other hand, increasing the number of rankings ($N = 30$) the BRE-MAXT looses its positive edge. This flaw of the BRE-MAXT performance can be explained due to the fact that a lot of quite similar rankings are included into the combination by the replacing process.
Since the 1T version outperforms significantly the NW schema in all the cases, the NW has been tested against the competitors only for the cases 1,4 ($N = 3$). Unfortunately, BRE-NW outperforms significantly median and Opt_list in terms of ρ and only in the case 1.

We point out that in a real situation the quality of the input rankings or their distribution is unknown. It can be quite difficult to determine if the rankings are heterogeneous or homogeneous in terms of quality with respect to the true ranking. This may introduce doubts about whether to apply BRE or other aggregation methods. Taking into account the average results among the three cases (*good, equal and poor* with$N = 10$, 30) showed in Tab. 3.2, we can assert that BRE-1T outperforms the competitors for both N values in terms of ρ and F. Only for $N = 30$ Opt_list shows a slightly better value F (.0825) of BRE-1T (.0987). We point out that the high results reported by Opt_list in terms of F are also related to the fact that Opt_list optimizes the F distance. For the results presented in Tab. 3.2, we conclude that BRE achieves better results than the competitors, and it can be used even if the distribution of the quality of the input rankings is not known *a priori*.

We can notice that BRE with the weighting schema (BRE-1T) gives a notable contribution to increase the performance with respect to the competitor methods in the cases where the quality of the rankings with respect to the true ranking is heterogeneous such as the cases *equal* and *poor*. In cases where the majority of the rankings are quite informative as in the case *good*, BRE provides also interesting results even if it outperforms significantly only the mean. As general consideration BRE can be applied successfully even if the quality of the ranking is not known *a priori* since BRE outperforms the competitors in average among all the three cases evaluated. The result of BRE in the cases *good* arises some interesting questions about the limits on BRE in case of identical rankings, this will be explored in Experiment 3 (Sec. 3.6).
From the results of BRE-MAXT, we conclude that the iterative schema increases the performance with respect to the 1T version with a not too big number of experts. This leads to the conclusion that using a simply computed number of iterations ($MAXT = \frac{N}{2}$) may not be the optimal solution for all the cases. The right number of iteration should take in consideration the distribution of the quality of the input rankings, in order to avoid the case where several similar rankings are introduced.

Table 3.3: Spearman correlation coefficient (ρ) and Spearman Footrule distance (F) of BRE using the mean and the raw mean as true-rank estimator. •means that BRE with mean is significantly better than BRE with raw mean as true-rank estimator

		Evaluation measure ρ									
Method	True-Rank	3 Rankers Cases				10 Rankers Cases			30 Rankers Cases		
	Est.	MAXT=3T				MAXT=5T			MAXT=15T		
		1	2	3	4	good	equal	poor	good	equal	poor
Mean		.4781	.5419	.7958	.0782	.9621	.8760	.7793	.9856	.9543	.8802
BRE-1T	Mean	.3826	.5742	.8226•	.0722	.9763•	.9207•	.8893•	.9903•	.9742•	.9353•
BRE-1T	rmean	.4322	.5717	.8214	.0733	.9751	.915o	.8751	.9896	.9686	.9176
BRE-MAXT	rmean	.3464	.5780	.8311	.0666	.9782	.9270•	.8880 •	.9785	.9714	.9372•
BRE-MAXT	rmean	.4237	.5786	.8314	.0701	.9780	.9250	.8750	.9800	.9725	.9273
		Evaluation measure F									
Mean		.4117	.4045	.2828	.5318	.1763	.2858	.3461	.1637	.2639	.3438
BRE-1T	Mean	.4938	.4132	.2578•	.6262	.0926•	.1688•	.1982•	.0592•	.0936•	.1482•
BRE-1T	rmean	.4697	.4147	.2591	.6282	.0953	.1749	.2115	.0614	.1036	.1672
BRE-MAXT	Mean	.5079	.4108	.2475	.6282	.0888	.1625•	.2014•	.0852	.1044	.1525•
BRE-MAXT	rmean	.4719	.4107	.2477	.6286	.0891	1648	.2131	.0842	.1023	.1600

3.5 Experiment 2: Raw Mean vs. Mean as Estimator

In this experiment we compare the performance of BRE where the true-rank estimator are the mean and the raw mean. The raw mean is the mean of the inputs rankings without the re-ranking step. As estimator the raw mean compared to the mean, presents an attractive feature that it has a lower computational cost. Some issues could arise from the fact that the raw mean does not produce a valid ranking and the F distance used for the weight computation is defined between valid rankings. We recall that the true-estimator method is a parameter of BRE, and the choice of a parameter should take into account the computational cost and the performance results. The goal of this experiment is to evaluate the performance of the raw mean as true-rank estimator, in order to use this estimator to decrease the computation time in the next experiments.

In order to assess empirically the performance of BRE with these two true-rank estimators, we use the synthetic data proposed in Experiment 1 (Sec. 3.4). The performance has been evaluated in terms of ρ and F. The statistical significance of the difference of the averages between BRE with mean and BRE with raw mean used as true-rank estimator is computed with a paired two-tailed t-Test (with $\alpha = 0.05$)on the 10 replicas.

From the comparison of the two estimators with BRE-1T Tab. 3.3, the mean outperforms significantly the raw mean in 7 cases over 10 (N=10, 30). With the regard to the MAXT version, the gap of the performance between mean and raw mean is reduced, since the mean produces better result than the raw mean in only in 3 cases over 10. As overall consideration, the difference of the BRE performance with raw mean instead of mean is not so dramatic even in the cases when the results are statistically significant. We point out that BRE with raw mean outperforms significantly the mean as competitor in the same cases highlighted in Tab. 3.1 using BRE with mean as true-rank estimator.

Although, the BRE with raw mean as estimator clearly does not outperform BRE with the mean, the use of the raw mean allows to decrease the computational cost and keeps valid the significantly performance of BRE against the mean in terms of ρ and F. For the motivation discussed above, in the next experiments we will use the raw mean instead of the mean as true-rank estimator in the BRE algorithm.

3.6 Experiment 3: BRE on Video Chunks Data in P2P network

In this section we describe the results of BRE, when the rankings to aggregate are highly correlated with the true ranking. The goal of this experiment is to evaluate and discuss the BRE performance with respect to the competitors in a more extreme setting where the input rankings are very similar among them and also highly correlated with respect to true rankings. The reason of this investigation arises from Experiment 1 (Sec. 3.4) where BRE have shown difficulties to aggregate rankings in the cases *equal* and *good* where the rankings show an homogeneous high quality.

For this experiment we use the data related to the transmission of video chunks packets in peer to peer (P2P) networks [49]. A video is transmitted into the network in several chunks of data and each peer receives data chunks in different order due to the presence of the delay. Detailed description of the problem and the simulator is provided in [49]. Ranking aggregation is not directly involved in the transmission problem but this data give us a chance to test our method on similar rankings also similar in terms of quality to the true ranking. Regarding our problem, the peers correspond to experts that produce rankings of the data chunks received and the true ranking is the right order of the chunks sent in the network. From this data we created two datasets as following:

C1000 : 100 peers with rankings composed of 1000 chunks. All the rankings show a ρ equal to .90 with respect to the true ranking,

C500 100 peers and rankings composed of 500 chunks. All the rankings show a ρ equal to .77 with respect to the true ranking.

The length of the ranking in input is increased than in the previous experiments ($n = 1000, 500$ instead of $n = 300$). Regarding to the number of rankings (N) to be aggregated even if there are available 100 rankers, we have decided to take into account $N = 3$ and $N = 40$. This decision is motivated by the fact that all the rankings are very identical and also to keep a number of rankings comparable with the Experiment 1 (Sec. 3.4). As in Experiment 1, the performance are evaluated in terms of Spearman correlation ρ and the Spearman footrule distance F. The statical significance has not been computed because there are no replicas in the datasets.

In this setting, mean and median are the most appropriate competitor methods, since we expect that the high number of identical rankings implies great performance from these two heuristic methods. In this experiment we have evaluated BRE-1T and BRE-MAXT versions, since they have showed the best performance in Experiment 1 (Sec. 3.4).

The mean and median show very good performance in terms of ρ and F in both datasets C1000 and C500. On the C1000 dataset, mean and median show the same performance

Table 3.4: Results of BRE with respect to the median and the mean on the video datasets. Performance are evaluated in terms of ρ and F distances with respect to the true ranking.

		Evaluation measure: ρ			
Method	True-Rank	C1000		C500	
	Est.	# peers		# peers	
		3	40	3	40
Mean		.9999	1	.8396	.8842
Median		.9999	1	.7817	.8126
BRE-1T	rmean	.9999	.9266	.8385	.8837
BRE-MAXT	rmean	.9999	.9354	.8075	.8830
		Evaluation measure: F			
Mean		4.4e-5	1.4e-4	.2425	.1835
Median		4.4e-5	1.4e-4	.3194	.2548
BRE-1T	rmean	4.4e-5	.1163	.2485	.2011
BRE-MAXT	rmean	1.12e-4	.1046	.2885	.2029

but in C500 where the rankings are still quite good the mean seems to overwhelm the median. On the C1000 dataset, BRE-1T and BRE-MAXT show performance equivalent to their true-rank estimators up to the approximation error. With 40 ranking BRE slightly decreases its performance. We notice the limitations of the weighting schema with respect to a strong heuristic methods such as the mean when the quality of the rankings is homogeneous. Same considerations are valid also for the C500 dataset, where BRE does not outperforms the mean in terms of ρ and F.

As we expect the results showed in this experiment highlight how BRE does not clearly outperforms the competitors evaluated in the cases of homogeneous quality rankings, BRE can at least tend to the performance of the mean used as an aggregation method. To conclude, we had also an empirical proof that the iterative process degrades the performance when several identical rankings are introduced in the combination.

3.7 QBRE: Quality Belief Ranking Estimator

In this section we describe Quality Belief Ranking Estimator (QBRE), an algorithm based on the BRE framework for the approximation of the quality of the rankings provided by the experts.

The approximation of the quality of the input rankings can be viewed as an interesting task in many real situations where there is limited a priori knowledge about the reliability of the experts involved in the combination. An example of this situation is the case of the miRNA target predictions [28]. miRNAs are small non coding RNA sequence that are involved in the protein regulation in animal and plants. miRNA sequences binds the mRNA sequences (messenger RNA), called target, to regulate the gene expression level of target or to degrade directly the mRNA target. Since single miRNA can bind several mRNAs and the validation of all possible targets throught experimental techniques is not yet feasible, so computational target predictor methods are the most useful sources to find putative miRNA targets. Since the growing number of miRNA target predictors

Algorithm 3 QBRE: Quality Belief Ranking Estimator

```
input I=$\tau^1, \ldots, \tau^N$ // a vector of N Rankings
input $S$ // Numbers of step
input $TE$ //  true-rank estimator method
input $\epsilon$ // Numeric precision
  k= 0
  BE=ComputeBelief_From_Rankings
  while  k != $S$  do
    $\bar{w}^k$=ComputeWeight(I,TE(I))
    BE=ApplyWeight($\bar{w}^k$,BE)
    FinalRank$_k$=Combination(BE)
    if  k $\geq$ 1 then
      TE= FinalRank$_k$
    end if
    if  $||\bar{w}^{k-1} - \bar{w}^k||_1 < \epsilon$ then
      break
    end if
    k++
  end while
output  $\bar{w}^k$
```

present in literature, the combination of the output of the miRNA target prediction is a valid technique to enhancement the prediction performance of the miRNA targets. On the other hand, the combination of the miRNA target predictions is a challenging problem due to the fact that there is a very small set of validated targets. In this scenario the quality and the reliability of each predictor method is quite hard to know *a priori* since in most of the cases these predictors are used by biologists as ”black box”. The weighting schema proposed in the BRE algorithm has shown to be effective and it can fit quite well the quality of the input rankings when a valid true estimator is used. The ranking output, produced by BRE, has shown great results with respect to the single true-estimator method tested. This results lead us to use the BRE's output as true-rank estimator in order to get weights that give a good approximation of the true weight of the rankers that are unknown in our unsupervised context.

As showed in Alg. 3, QBRE is based on the same main components described for BRE (Sec. 3.2) such as the mapping of the rank into bba, the combination of the bba and the weights computation with the relative bba discount. QBRE differs from BRE in the output in fact QBRE returns a vector of weights $\bar{w}^k$ instead of a ranking. BQRE needs also as input a numerical constraint ϵ used to check the convergence to a stable solution. QBRE has input also the true-rank estimator method (TE), that is used in the first step to produce the combined ranking. The core difference of BQRE against BRE is the iterative step. In the first iteration, QBRE works as BRE and it produced a combined ranking (using the weights computed by $TE(I)$). At each step the current final ranking at step k-th is used as true-rank estimator for the $k+1$-th step. The algorithm returns the weights $\bar{w}^k$ as output when the p-norm with $p = 1$ ($||\bar{w}^{k-1} - \bar{w}^k||_1$) is less than ϵ. The combined presence of the number of steps S in the while condition and the ϵ as the stop criteria, will lead or eventually force the algorithm to output a vector of weights.

Moreover, ϵ gives the possibility to specify a fixed precision value required in the results.

3.8 Experiment 4: Evaluation of QBRE

In this section we present the evaluation of QBRE on the task of the quality ranking approximation. To the best of our knowledge we have not found other competitor methods of QBRE for the specific problem at hand. Moreover, we are interested to assess if the weights provided by QBRE are qualitatively better than the weights provided by the BRE algorithm. The expectation of this experiment is to find better weights with respect to the weights found by BRE.

We evaluated the quality of the weights founded by QBRE against the weights produced by BRE on the same synthetic data generated for Experiment 1 (Sec. 3.4). In order to measure the quality of the weights we use the absolute and relative error with respect to the true weights computed from the true ranking. The absolute and the relative error of the weights are defined as:

$$E(\bar{\mathbf{w}}, \bar{\mathbf{w}}^*) = \frac{\sum_{j=1}^{N} |w_j - w_j^*|}{N}$$

$$RE(\bar{\mathbf{w}}, \bar{\mathbf{w}}^*) = \frac{\sum_{j=1}^{N} \alpha_j * |w_i - w_j^*|}{N} \qquad \alpha_j = \frac{1}{w_j^*}$$

where N is the number of the rankers and $\bar{\mathbf{w}}$,$\bar{\mathbf{w}}^*$ are respectively the vector of the weights produced by the evaluated method using the Eq. 3.3 and the vector of the true weights. The true weight w_j^* for the j-th ranking is computed as $w_j^* = F(\tau^j, \tau^{T_{rank}})$ where F is the Spearman Footrule distance (Eq. 2.1). With regard to the relative error RE, α_i coefficient represents the importance of the j-th ranking. For both the errors used, a better weight approximation corresponds to a smoother value. In Tab 3.5 are shown the results of the comparison between BRE-1T and BQRE in terms of E and RE errors. As in Experiment 1 (Sec. 3.4), the reported results are the average along the 10 replicas, and also the statistical significance is computed in the same way (a paired two-tailed t-Test with $\alpha = 0.05$). As true-rank estimator, we have used the *raw mean* (rmean), that corresponds to the mean of rankings without the re-ranking step.

With regard to QBRE, we set 10 as max number of steps and $\epsilon = 0.5e - 4$. To have a fair comparison we have also used the raw mean inside the QBRE as estimator.

From Tab. 3.5, we observe that the weights produced by QBRE improve significantly BRE-1T in both error measures. Even if we do not show the numerical stability of the weights found, QBRE obtains stable weights for a fixed ϵ in less of 5 step in all the cases.

In this experiment we have presented and evaluated the QBRE algorithm which aims to provide an effective estimation of the quality of the rankings with respect to the unknown

Table 3.5: Absolute and relative errors between the weights provided by BRE-1T and QBRE with respect to the true weights. The statistical significance of the QBRE results with respect to BRE-1T is denoted by •.

			E error							
	3 Ranking cases				10 Rankings cases			30 Rakings cases		
Method	1	2	3	4	good	equal	poor	good	equal	poor
BRE-1T	.2211	.1399	.0936	.2662	.0651	.0830	.1341	.0625	.0763	.1067
QBRE	.2096	.1075•	.0533•	.2222•	.0158•	.0201•	.0463•	.0139•	.0133•	.0188•
				RE error						
BRE-1T	.4698	.2480	.2358	.4154	.3432	.2414	.546	.3449	.3106	.2859
QBRE	.5000	.2214•	.1549•	.3488•	.0989•	.0541•	.1901•	.0887•	.0649•	.0559•

true ranking. On the limited cases evaluated on the synthetic data, QBRE has shown to be an effective and simple algorithm that can provide a good estimation of the quality of the input rankings. As future work it could be interesting to explore QBRE on a real environment where QBRE's output could help the user to give a rough quality evaluation of the experts.

3.9 About The Weighting Schema

As described in Sec. 3.8, BRE algorithm has two key points that could be modified to increase its performance for the estimation of the true ranking. The first point is the quality of the weights, and this was investigated in Experiment 4 (Sec. 3.8). The other point is the application of the weights in terms of weighting schemas. As weighting schemas we mean different criteria for which select the good and the poor rankings from the weights retrieved. The aim of different weighting schemas applied inside BRE is to model the impact of the good rankings when the uncertainty on Θ is transferred to P, and on the other hand the impact of the poor rankings when belief on P is transfered to Θ. Also the application of the weights on the bba's defined on Θ could be also modified, (for example changing the belief on $\neg P$) but we have decided to left it unchanged as previously defined in Eq. 3.4, since no evidence on $\neg P$ is introduced.
We recall the base weighting schema introduced in Sec. 3.2:

$$\begin{array}{ll} if \quad w_j = min(\{w_1,\dots,w_N\}) & if \quad w_j \neq min(\{w_1,\dots,w_N\}) \\ m'_{ji}(P) = m_{ji}(P) + (w_j * m_{ji}(\Theta)) & m'_{ji}(\Theta) = m_{ji}(\Theta) + (w_j * m_{ji}(P)) \\ m'_{ji}(\neg P) = 0 & m'_{ji}(\neg P) = 0 \\ m'_{ji}(\Theta) = 1 - m'_{ji}(P) & m'_{ji}(P) = 1 - m'_{ji}(\Theta) \end{array} \tag{3.6}$$

where $j \in 1,\dots,N$ and $i \in 1,\dots,n$ indicate respectively the rankings and the items. We refer to the above weighting schema as *base schema*, that we have used in all the previous experiments up to now (Experiment 1, 2 and 3 Sec. 3.4, 3.5, 3.6). In the base schema, only for the items of the most informative ranking ($w_j = min(\{w_1,\dots,w_N\})$) the belief of P is increased whereas the item's bba of the less informative rankings

($w_j \neq min(\{w_1, \ldots, w_N\})$) are modified in the opposite sense. A possible drawback of the base schema is that there is no difference among all the not informative rankings because for all of them the mass on P is transferred to Θ.

The base schema singles out the most informative ranking from the others quite well but we want to evaluate a more smooth criterion to weight the highly-informative rankings. To do that we have decided to use the linear deviation from the mean of the weights in order to select more rankings as highly-informative rankings, defining the **version 0** schema as:

$$\begin{array}{ll}
if \quad d_j \geq 0 & if \quad d_j \leq 0 \\
m'_{ji}(P) = m_{ji}(P) + (w_j * m_{ji}(\Theta)) & m'_{ji}(\Theta) = m_{ji}(\Theta) + (w_j * m_{ji}(P)) \\
m'_{ji}(\neg P) = 0 & m'_{ji}(\neg P) = 0 \\
m'_{ji}(\Theta) = 1 - m'_{ji}(P) & m'_{ji}(P) = 1 - m'_{ji}(\Theta)
\end{array}$$

where d_j are the linear deviation from the weight rankings to the mean defined as $d_j = \frac{\sum_{k=1}^{N} w_k}{N} - w_k$. All the rankings with $d_j \geq 0$ are classified as informative rankings and their $m_{ij}(P)$ will be increased according to left the part of Eq. 3.6. On the other hand, rankings with $d_j \leq 0$ are classified as not informative rankings, and consequentially their m_{ij} on P will be decreased (right side in Eq. 3.6). We point out that in all the versions, the w_j values are the weights computed as in Eq. 3.3. Starting from the version 0 schema we propose the following other three schemas:

Version 1

$$\begin{array}{ll}
if \quad d_j \geq 0 & if \quad d_j \leq 0 \\
m'_{ji}(P) = m_{ji}(P) + (d_j * m_{ji}(\Theta)) & m'_{ji}(\Theta) = m_{ji}(\Theta) + (w_j * m_{ji}(P)) \\
m'_{ji}(\neg P) = 0 & m'_{ji}(\neg P) = 0 \\
m'_{ji}(\Theta) = 1 - m'_{ji}(P) & m'_{ji}(P) = 1 - m'_{ji}(\Theta)
\end{array}$$

Version 2

$$\begin{array}{ll}
if \quad d_j \geq 0 & if \quad d_j \leq 0 \\
m'_{ji}(P) = m_{ji}(P) + ((1 - w_j) * m_{ji}(\Theta)) & m'_{ji}(\Theta) = m_{ji}(\Theta) + (w_j * m_{ji}(P)) \\
m'_{ji}(\neg P) = 0 & m'_{ji}(\neg P) = 0 \\
m'_{ji}(\Theta) = 1 - m'_{ji}(P) & m'_{ji}(P) = 1 - m'_{ji}(\Theta)
\end{array}$$

Version 3

$$\begin{array}{ll}
if \quad d_j \geq 0 & if \quad d_j \leq 0 \\
m'_{ji}(P) = m_{ji}(P) + ((1 - w_j) * m_{ji}(\Theta)) & m'_{ji}(\Theta) = m_{ji}(\Theta) + ((1 - w_j) * m_{ji}(P)) \\
m'_{ji}(\neg P) = 0 & m'_{ji}(\neg P) = 0 \\
m'_{ji}(\Theta) = 1 - m'_{ji}(P) & m'_{ji}(P) = 1 - m'_{ji}(\Theta)
\end{array}$$

Table 3.6: Results of BRE-1T with the different weighting schemas on the synthetic data. Performance are evaluated in terms of ρ and F distances with respect to the true ranking. The statistical significance of the base weighting schema with respect to all the other schema is denoted by •, instead of ∘that means the opposite case.

			Evaluation measure ρ									
True-Rank	Method	Weight.	3 Rankers Cases				10 Rankers Cases			30 Rankers Cases		
Est.		Schema	1	2	3	4	good	equal	poor	good	equal	poor
rmean	BRE-1T	base	.4322	.5717∘	.8214	.0733	.9751•	.915	.8751	.9896•	.9686•	.9176
rmean	BRE-1T	v0	.4690	.6137	.81642	.0778	.9705	.9129	.8512	.9629	.9640	.9203
rmean	BRE-1T	v1	.4690	.6137	.8164	.0778	.9705	.9129	.8512	.9772	.9647	.9200
rmean	BRE-1T	v2	.4690	.61377	.8164	.0778	.9707	.9129	.8512	.4505	.9339	.9198
rmean	BRE-1T	v3	.4342	.6226	.8172	.0738	.9699	.9138	.8588	.4954	.9435	.9216
			Evaluation measure F									
rmean	BRE-1T	base	.4690	.4147∘	.2591	.6282	.0953•	.1749	.2115	.0614•	.1036•	.1672
rmean	BRE-1T	v0	.4534	.3960	.2641	.6245	.1020	.1771	.2316	.1121	.1122	.1635
rmean	BRE-1T	v1	.4534	.3960	.2641	.6245	.1020	.1771	.2316	.0911	.1087	.1630
rmean	BRE-1T	v2	.4534	.3960	.2641	.6245	.1021	.1772	.2316	.4723	.1517	.1657
rmean	BRE-1T	v3	.4641	.3923	.2615	.6264	.1038	.1763	.2256	.4443	.1411	.1633

In the base version (Eq. 3.6), the weights for the most informative rankings are near to 0, since lower values of F means rankings more similar to the ranking produced by the true-rank estimator. In the version 1 and version 2 respectively we evaluate lower and higher weight values for the most informative rankings. The rationale underlying the use of $1 - w_j$ as weight in the version 2 is based on the idea to evaluate how the rankings are similar with respect to the inverse of the true-rank estimator. Moreover, in the version 3, also the weights for the low informative rankings are modified to evaluate the effect of the $1 - w_j$ values in both cases.

3.10 Experiment 5: Evaluation of The Weighting Schemas

In this experiment we will evaluate the performance of BRE with the proposed weighting schemas, in order to assess which is the best weighting schema for the BRE algorithm. We have used the synthetic data of Experiment 1 (Sec. 3.4). The performance are measured like in the previous experiments in terms of Spearman correlation ρ and Spearman footrule distance F. The statistical significance of the differences of the performance of the weighting schemas is computed with a paired two-tailed t-Test (with $\alpha = 0.05$) on the 10 replicas. We evaluated only the BRE-1T version, due to the fact that the weighting schema effect is clearly visible at the first iteration.

From Tab. 3.6 we notice that the four weighting schemas do not improve uniformly the base schema. In the case *good* (N=10, 30) and in the case *equal* ($N = 30$) the base solution outperforms significantly all the other weighting schema. With N=3 the base solution shows competitive performance with respect to the weighting schema evaluated, but only in case 2 ($N = 3$) the base weight is significantly worst than the by the other schemas proposed. Moreover, it seems that there is no difference among the version 0, 1 and 2. With an high number of rankings (N = 10, 30), all versions proposed degrade significantly the performance with respect to the base version. This fact can be explained as a sort of saturation of the combined masses due to our belief assignments over the frame Θ. For all the ranking items we assign simple belief function always on P, so when several rankings increase their bbas on P (as in weighting schema version 0,1 and 2) this brings to 1 the combined belief on P for many items producing ties in the output rankings. To avoid this fact an assignment on Θ that takes into account also the belief on $\neg P$ should be proposed, but this is beyond the empirical evaluation of BRE discussed in this chapter. This option will be faced in future work. We conclude that the base schema is the best weighting schema with respect to the other schemas proposed in this experiment. The cause of the better performance of the base version lies in its simplicity and probably it avoids the problem of saturation mentioned above.

3.11 Experiment 6: BRE with QBRE Weights vs. BRE

In the latter section we explored the two key points of the proposed BRE algorithm: the role of the weights (Sec. 3.7-3.8), and the impact of different weighting schemas (Sec. 3.9-3.10). We would conclude our explorative work, with an experiment that aims to verify the quality of the BRE algorithm when the best weights according to QBRE are used.

In this experiment we have compared the weighted version of BRE against a version of BRE where the weights are provided by the QBRE algorithm (referred as BRE-1T (QBRE-weights)). To have a fair comparison we have to evaluate the 1T version since the input rankings are the same only in the first iteration. As true-rank estimator we have used the raw mean in both methods BRE-1T and QBRE. The results showed in Tab. 3.7, include also the BRE-NW and the iterative version of BRE (BRE-MAXT) in order to show an exhaustive comparison of all the BRE versions discussed. As in the previous experiments we have used the synthetic data described in Sec. 3.4. The performance are evaluated using the Spearman correlation coefficient ρ and Spearman footrule distance F and also the statistical significance is computed in the same way, namely using a paired two-tailed t-Test with $\alpha = .05$.

As in our expectation BRE with the weights founded by QBRE performs significantly better in all the cases except for case 1 and 4 with $N = 3$. We point out that BRE 1T with the weights found by QBRE outperforms significantly also the BRE-MAXT in all the three cases with $N = 30$. This results shows again the limits of the BRE-MAXT with an high number of rankings. As discussed in Experiment 1 (Sec. 3.4). The superior performance of BRE-1T with QBRE weights is remarkable in the cases *poor* (N = 10, 30) where the QBRE weights make the difference with respect to the use weights computed through the true-rank estimator.

From Tab. 3.7 according to conclusion of the previous experiment (Sec. 3.10), we have also highlighted the best results of the global weights where the weight of each ranking is discounted to all its items. For a further improvement of the performance, a solution based on local weights associated to each item should be explored. The weight of the single items or a subset of items can still be computed by the true-rank estimator or in addition with *a priori* knowledge on the items. For example if we know that some experts can express a wrong ranking on particular items, this can be modeled with local weights defined appropriately for each rankings. Our intuition on local weights inside BRE, is the possibly to avoid the ties problem mentioned in the previous experiments.

Table 3.7: BRE with QBRE Weights vs BRE: Average ρ and F distance of BRE-1T with the weights evaluated on the synthethic data. The statistical significance of BRE -1T (QBRE Weights) with respect to BRE -1T is denoted as •, instead of ∘ that means the opposite case.

Method	True-Rank. Est.	3 Rankers Cases MAXT=3T				10 Rankers Cases MAXT=5T			30 Rankers Cases MAXT=15T		
		1	2	3	4	good	equal	poor	good	equal	poor
Mean		.4781	.5419	.7958	.0782	.9621	.8760	.7793	.9856	.9543	.8802
BRE-NW		.4888	.5254	.7799	.0804	.9383	.8453	.7723	.9409	.8941	.8074
BRE-1T	rmean	.4322	.5717	.8214	.0733	.9751	.915	.8751	.9896	.9686	.9176
BRE-MAXT	rmean	.4237	.5786	.8314	.0701	.9780	.9250	.8750	.9800	.9725	.9273
BRE-1T	QBRE weights	.3985	.5736	.8250•	.0682∘	.9768•	.9239•	.8981•	.9903•	.9753•	.9704•
Evaluation measure F											
Mean		.4117	.4045	.2828	.5318	.1763	.2858	.3461	.1637	.2639	.3438
BRE-NW		.4511	.4393	.2912	.6235	.1444	.2368	.3129	.1444	.1919	.2669
BRE-1T	rmean	.4697	.4147	.2591	.6282	.0953	.1749	.2115	.0614	.1036	.1672
BRE-MAXT	rmean	.4719	.4107	.2477	.6286	.0891	.1649	.213	.0824	.1023	.1600
BRE-1T	QBRE weights	.4800	.4127	.2550 •	.6299	.0917•	.1655 •	.1882 •	.0593 •	.0917 •	.1419 •

This experiment concludes the experimental work on BRE algorithm began in Experiment 4 (Sec. 3.8), that aims to explore the role of weights and their application inside the algorithm. On the synthetic data we have showed how the BRE-1T version can effectively provide quite good true ranking approximation, when the weights are good approximation of the quality of the input ranking. However, this consideration is not evaluated on real data and not all the true-rank estimators have been evaluated.

3.12 Conclusions

In this chapter we presented Belief Ranking Estimator (BRE), an unsupervised method that estimates a true ranking given a set of estimating ranked permutations. BRE, through the use of the belief function framework, models the uncertainty of each ranking and combine them accordingly to the weights computed as distances from a true-rank estimator that can be provided by any ranking aggregation method. In this chapter we focused on the evaluation of BRE in the case of total rankings.

From results on synthetic data, with low-quality input rankings BRE with base weighting schema has provided better estimation of the true ranking with respect to the mean, median and the Footrule optimal aggregation method used as competitors. Moreover, we point out that BRE shows significant performance with respect to the competitors also when an increasing number of rankings is involved. The BRE algorithm has not shown so brilliant results when all the combined rankings are the same, whereas mean and median achieve better performance. We explored empirically two main aspects of BRE: the quality of weights, and the weighting schemas used. With regard to the quality of weights we have presented a novel algorithm based on BRE, called Quality BRE (QBRE), that aims to approximate the true weights of the rankings involved in the combination. QBRE has provided qualitative better weights with respect to BRE. On the other hand, several different weighting schemas has been evaluated on BRE, but the base weighting schema has shown the best results. Finally BRE with weights computed by QBRE and BRE has been compared, showing that BRE with good quality weights improve the performance significantly.

With regard to the total rankings, some open issues of BRE algorithm should be investigated in future works. Due to the difficulties of BRE to combine similar rankings, a procedure that discovers from the input data when the rankings show enough heterogeneous quality could suggest the use of BRE or not. The number of iterations in the BRE iterative version is another issue not investigated, a method that finds the optimal iteration numbers will be an interesting mean to increase the performance. Instead of using global weights that measure the quality of the rankings, local weights devised for each item can give the possibility to manage *a priori* partial/total knowledge on the items. This consideration will imply also the exploration of different belief assignments over the frame Θ.
Among all the open issues listed, we will focus in the next chapter on the investigation of BRE on partial rankings, due to the fact that partial rankings are met in most of the real problems where rankings are involved.

Chapter 4

Belief Ranking Estimation Applied to Partial Rankings

4.1 Aggregation of Partial Rankings and Top-k lists: Definition

Let U be a set of n items on whose subsets N experts produce rankings. We denote as S_j with $S_j \subseteq U$ the set of items given in input to the j-th expert, S_j has cardinality $s_j = |S_j|$. Each expert produces a ranking $\tau^j = (\tau^j(1), \ldots, \tau^j(i), \ldots, \tau^j(l_j))$ where $\tau^j(i)$ is the rank associated to the item $x_i \in C_j$ and C_j is the set of items contained in the ranking. Moreover, we denote as $l_j = |\tau_j|$ the length of the j-th ranking, namely the number of the items ranked. We suppose to have $\tau^{T_{rank}} = (\tau^{T_{rank}}(1), \ldots, \tau^{T_{rank}}(n))$, that is the golden true ranking, namely a total rank on the set U. Depending on the items ranked in τ three possible cases arise:

Total Rankings Total rankings are rankings that contain the same set of items, all the experts have in input exactly the entire set U ($\forall j \in 1 \ldots N, S_j = U$ and $\forall j \in 1, \ldots, N, |\tau_j| = n$). Total rankings, namely permutations, have been widely discussed in Chapter 3.

Partial Rankings Partial rankings occur when the rankings are induced by a total ordering over the set of items S_j. Our simplifying assumption is that $l_j = |S_j|$, namely the length of the ranking corresponds to the cardinality of entire set of items of the expert, so for each expert we have that $l_j = |\tau_j| < |U|$. In this case if an item is not present in the ranking we assume that it does not belong to the items ranked by the expert knowledge.

Top-k Rankings/list For each ranking τ_j only the corresponding top k_j items are included in the ranking so $l_j = |\tau^j|$ with $k_j = l_j$. In other words, only a subset of S_j is included in the ranking. In this case if an item is not present in the ranking we are not sure if it belongs or not to the set S_j of the expert.

We point out that the set S_j, namely the input set of items of each expert, is totally unknown in real problems. Moreover it is quite hard to have any partial knowledge of S_j since as input we have only rankings of different length l_j that can contain totally different items (the items ranked are denoted as C_j). The notion of S_j even if unknown

Algorithm 4 Belief Ranking Estimator: Iterative version for partial/top-k rankings

```
input I=$\tau^1, \ldots, \tau^N$ // a vector of N partial rankings
input $T$ // Numbers of iterations
input $TE$ // True-rank estimator
  I=Augmented_Rankings(I)
  k= 0
  BE=Belief_From_Rankings(I)
  $FinalRank_k$=Combination(BE)
  while k != $T$ do
    W=ComputeWeights(I,TE(I))
    BE=ApplyWeights(W,BE)
    $FinalRank_k$=Combination(BE)
    I[pos(max(W))]=$FinalRank_k$
    BE=Belief_From_Rankings(I)
    k++
  end while
output $FinalRank_k$
```

has been introduced to explain better the differences between partial and top-k rankings. We know that the assumption $l_j = |S_j|$ is quite a strong and limiting constraint, since we can still have a partial ranking also for a subset of S_j. For the number of possible hypotheses that can be formulated for the partial rankings is big to explore, we have focused on this assumption, in order to give a sufficient evaluation of BRE with synthetic data. The problem in its general form is stated as follows.
Given N partial or top-k rankings τ^j that estimate with unknown quality the unknown true ranking $\tau^{T_{rank}}$ find a ranking that estimates the true ranking.

To measure the disjunction of the set of items contained in the input rankings, we introduce the $DisJ$ coefficient. Given a set of input partial/top-k ranking $\tau^1, \ldots, \tau^j, \ldots, \tau^N$ where $U^* = \bigcup_{j=1 \ldots N} C_j$, the $DisJ$ coefficient is defined as:

$$DisJ = \frac{|U^*|}{N * k} \tag{4.1}$$

where U^* is the union set of all items ranked in the input rankings. The $DisJ$ coefficient is equal to 1 when all the items are different (totally disjointed) and it is equal to $\frac{1}{N}$ when all the rankings have the same items (total rankings). In the case of total disjunction $DisJ = 1$, the task of the estimation of the true ranking increases its difficulty. In particular BRE faces a possible absence of one or more belief function assignment to the some items.

4.2 BRE applied to Partial Rankings

The BRE algorithm previously described in Sec. 3.2, estimates the true ranking given a set of rankings. In Alg. 4 is showed the BRE algorithm for the partial rankings, that is substantially the same pseudo code presented for the total rankings. The input parameters are still the input rankings τ^j, the number of iterations T and the true-rank estimator method (TE). The main steps of BRE deeply described in Sec. 3.2 remain the same: the mapping of the item ranks into bba (*Belief_From_Ranking*), the weights computation from the true-rank estimator used (*ComputeWeights*), the application of the weights to the current belief model of the rankers (*ApplyWeights*), finally the output ranking is produced by the combination of all the bbas of the rankings. The main differences introduced in BRE to deal with partial rankings are:

- Trasformation of the partial/top-k rankings into special rankings called augmented rankings [5].

- The belief assigment of each item contained in the augmented rankings.

In the following sections, we will discuss in detail the modifications introduced in the BRE method in order to apply it to aggregate the partial rankings.

4.2.1 From Partial/Top-k Rankings to Augmented Rankings

We have to recall that BRE works basically on total rankings. This means that the set of items is the same for all the input rankings, in fact in the combination step BRE has a belief assignment over Θ for all the items. In the case of top-k and partial rankings, the rankings could be different in terms of items ranked and length, this opens the issue to transform the partial/top-k rankings into rankings that have the same set of items. We highlight the fact that we do not know *a priori* if a ranking is partial or a top-k, so we treat them in the same way. This issue has been resolved, introducing the augmented rankings [5]. Let τ^j be a ranking of length $k_j = |\tau^j|$, the augmented ranking τ^{*j} is defined as follows:

$$\tau^{*j}(i) \begin{cases} \tau^j(i) & if \quad x_i \in C_j \\ k_j + 1 & if \quad x_i \in U^*/C_j \end{cases} \tag{4.2}$$

where C_j is the set of ranked items in the τ^j ranking and $U^* = \bigcup_{j=1 \ldots N} C_j$ is the union of the items ranked in all the rankings inputs. U^*/C_j denotes the set-theoretic difference between the two sets, and it includes the items not ranked in the ranking τ^j. This operation is done in the *Augmented_Ranking* routine in Alg. 4, after this pre-processing step the input rankings consist in N augmented rankings τ^{*j} with length $|\tau^{*j}| = |U^*|$. An augmented ranking has the same rank values of the items as in the τ^j except for the items $x_i \in U^*/C_j$ that are added below all the original items with a rank value equal to $k_j + 1$. The idea to associate a rank value of $k_j + 1$ to all the items not present in τ^j models the fact that not having enough information to decide the right position of these items, we put them at the same position that is just after the last ranked item. Using the augmented rankings we solve the problem of having rankings of different length on heterogeneous set of items, in fact we obtain input rankings similar to total rankings with ties.

4.2.2 From Augmented Ranking to bba's

The frame of discerment $\Theta = \{P, \neg P\}$ is the same used in the case of total rankings. In the case of partial/top-k rankings the belief assigned on $\Theta = \{P, \neg P\}$ of each item should take into account the information of the added items of the aumented rankings. Given a set of N augmented rankings $\tau^{*1}, \ldots \tau^{*j}, \ldots, \tau^{*N}$ of length $|\tau^{*j}| = |U^*| \quad \forall j \in 1, \ldots, N$, the bba of the j-th ranking on the i-th items is consequently assigned as follows:

$$\begin{array}{ll} if \quad x_i \in C_J & if \quad x_i \in U^*/C_j \\ m_{ji}(P) = \dfrac{k_j - (\tau^{*j}(i) - 1)}{k_j} & m_{ji}(P) = 0 \\ m_{ji}(\neg P) = 0 & m_{ji}(\neg P) = 0 \\ m_{ji}(\Theta) = 1 - \dfrac{k_j - (\tau^{*j}(i) - 1)}{k_j} & m_{ji}(\Theta) = 1 \end{array} \tag{4.3}$$

where C_j is the set composed of the k_j items x_i contained in the original rankings τ^j.For all the items $x_i \in C_J$, the assignments over P and Θ are the same used for the total rankings (Eq. 3.2). The bba definition reflects the fact that highly relevant elements should have more belief to be in the right position. For the items added by the augmented rankings at position $k_j + 1$, the bba gives all the belief to Θ, due to the fact that we do not have any information about the correctness of the position of these items. The bba assignment for the items $x_i \in U^*/C_j$ is the vacuous belief function, and it represents the total ignorance over the possible hypothesis of the frame Θ. The vacuous belief function is also the neutral element in the conjunctive combination rule (Eq. 2.9) used inside BRE. In this way the belief function associated to the items $x_i \in U^*/C_j$ does not give any contribution in the combination and consequently only the belief functions related to the items $x_i \in C_j$ contribute to the conjunctive rules. As for the total rankings, the bba may reflect some *a priori* knowledge about the correctness of the items, other bbas derived for the specific problem has been evaluated on the LETOR datasets (Sec. 4.7).

4.2.3 Weight Computation and Weighting Schema

As for the total rankings, the weights are computed as distances between the input rankings to the ranking provided by the true-rank estimator (TE). In this case the ranking produced by the true-rank estimator is a total ranking over the set U^* and the input rankings are augmented rankings that are not properly total rankings. We have still used the Spearman footrule distance [4] as follows:

$$w_j = \frac{F(\tau^{*j}, \tau^{TE})}{\frac{1}{2}|U^*|^2} \quad \forall j \in 1..N \tag{4.4}$$

where τ^{TE} is the ranking produced by the true-rank estimator and $F(\cdot\,,\cdot\,)$ is the Spearman footrule distance defined over two total rankings. Although there are more appropriate distances between partial rankings and total ranking, we decided to use the footrule distance for its simplicity as base version. Other distances specifically designed to measure the similarity between partial and total rankings such as the induced and the scaled Spearman footrule distance [4](Sec. 2.2) has been also evaluated on the LETOR datasets.

The combination step remains the same described in Eq. 3.5, based on the conjunctive combination rule of the bba of each item among all the rankings.
With respect to application of the weights to the bba's, we have applied to the same formula used for the total rankings that we recall as follows:

$$\begin{array}{ll} if \quad w_j = min(\{w_1, \dots, w_N\}) & if \quad w_j \neq min(\{w_1, \dots, w_N\}) \\ m'_{ji}(P) = m_{ji}(P) + (w_j * m_{ji}(\Theta)) & m'_{ji}(\Theta) = m_{ji}(\Theta) + (w_j * m_{ji}(P)) \\ m'_{ji}(\neg P) = 0 & m'_{ji}(\neg P) = 0 \\ m'_{ji}(\Theta) = 1 - m'_{ji}(P) & m'_{ji}(P) = 1 - m'_{ji}(\Theta) \end{array} \tag{4.5}$$

where m_{ji} is the bba of the j-th ranking on the i-th item, $Mmin(\cdot)$ is the minimum function and m'_{ji} is the modified one. Also for partial rankings, the weights are applied globally to all the bba items of each expert. For the ranking that increases the belief on P ($w_j = Min(\{w_1, \dots, w_N\})$), also the vacuous belief function assigned to the augmented items will be interested. In this way unknown items for a ranking increase their belief on P due to the fact that the weights are applied indistinctly to the real and the augmented items.

4.2.4 The Iterative step

About the iterative schema, the worst ranking is replaced by the combined one as in the total rankings case, and the maximum number of iteration has been fixed as $MAXT = \frac{N}{2}$. We have to point out that the combined ranking produced is a total ranking over the set U^*, since we order the $BetP(P)_i$ for all the items $x_i \in U^*$. Inside the iteration, we replace the augmented rankings in input with total rankings and at the iteration $T = MAXT$ the half of the rankings will be total rankings on the U^* sets. The final ranking, denoted as O provided by BRE is of length $|U^*| = |O|$, but it could be also transformed in a top-k list if the user specify a valued for k.

Also for partial rankings we will evaluate both the not-weighted version ($BRE - NW$) and the iterative one. In the remaining of the chapter we refer as weighting schema to the $BRE - 1T$ version, whereas as iterative schema to BRE when $T = MAXT$.

Table 4.1: Average *DisJ* coefficients on the 10 replicas for all the cases generated and the different k values. The lowest value ($\frac{1}{N}$) for the values of $N = 3, 10, 30$ are respectively .3, .1, .03.

	N=3 min DisJ=.30				N=10 minDisJ=.10			N=30 min DisJ=.03		
k	1	2	3	4	good	equal	poor	good	equal	poor
15	.9356	.9044	.7311	.9461	4033	.5373	.674	.3009	.3636	.4078
60	.8006	.7761	6772	.8156	.3105	.3733	.41	.1457	.1548	.1591
120	.6494	.638	.5858	.6526	.2204	.2399	.2461	.0822	.083	.0832
210	.4635	.4568	.4444	.4591	.1414	.1428	.1429	.0476	.0476	.0476

4.3 Experiment 1: BRE on Top-k Lists

In this experiment the goal is to evaluate BRE against the other aggregation methods in the case of top-k lists with the same length k. As previously discussed, we have decided to use synthetic data on BRE, in order to have a complete control on the generation parameters of the rankings investigated.

As competitor methods based on heuristic methods we have included the mean and median of the rankings. As optimized method we have included a method, denoted as *AggrList*, that approximates the Footrule optimal aggregation using the Monte Carlo cross-entropy approach [14]. We have used the implementation provided by the R package *RankAggr* with the base parameters suggested by the authors. Even if the *Opt_List*, used in the total rankings, and *AggrList* are based on the same minimization problem we decided to refer to the latter one with a different name to mark the fact that it is an approximate solution of the minimization problem. The Markov chain based solution has not been included for the same motivations discussed for the total rankings Sec. 3.3.

In this experiment we evaluate the combinations of N top-k τ^j rankings where $j \in 1..N$, $n = |U| = |S_j|$ and all the rankings have the length $k = |\tau^j|$. All the experts have in input all the universe set U but the rankings outputted are limited to the top-k items. The generation of the data and the different quality cases evaluated are based on the same generation criteria adopted for the total rankings (3.4). We have fixed as $\tau^{T_{rank}}$ a total ranking of 300 items (n=300), and from it we have generated randomly the permuted ranking accordingly to different Spearman correlation coefficient ρ [20][22] values. As for the total rankings we have set the number of rankings N equal to $3, 10, 30$ and the same 10 different quality cases described in (3.4) has been evaluated. The difference from the previous generation is that we select the top k items from the permuted ranking. For all the 10 cases, we have evaluated k equal to $15\,, 60\,, 120\,, 210$, that correspond to $5\%\,, 20\%\,, 40\%\,, 70\%$ of the length n of the total ranking. Moreover for each case and for each N value we performed 10 independent replicas of the procedure using the same generation parameters in order to evaluate the statistical significance of the results. As in the previous experiments we have used a paired two-tailed t-Test with $\alpha = 0.05$. We point out that the statistical significance of the result of BRE with respect to the mean is indicated in the Tab. 4.2. For all the other competitors the significance of the result is indicated in the discussion. In order to evaluate the partiality of input rankings gen-

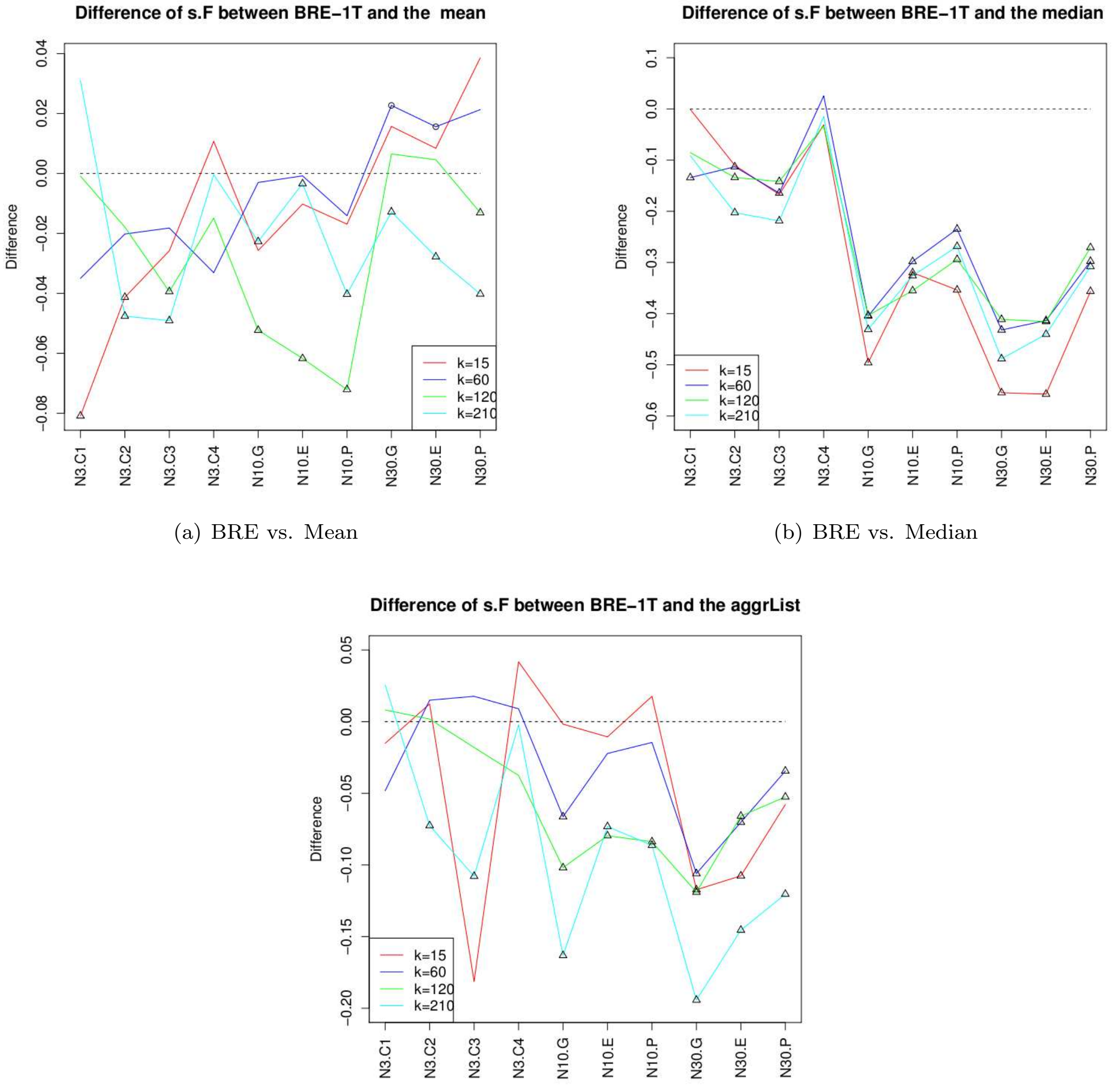
Table 4.2: Top-k rankings: Average of the scaled Spearman footrule distance ($s.F$) of BRE and of the competitor methods with respect to the true ranking. ● means that BRE is significantly better than the mean, and ○ means that BRE is significantly worse.

		Evaluation measure: $s.F$									
k	Method	N=3				N=10			N=30		
	cases	1	2	3	4	good	equal	poor	good	equal	poor
15	Mean	.6382	.5853	.4169	.648	.1644	.2538	.3582	.0910	.1400	.1898
	Median	.5582	.6542	.5582	.6898	.6347	.5636	.6951	.6613	.7058	.5849
	AggrList	.5724	.5316	.5724	.6169	.1404	.2542	.3236	.2240	.2560	.2862
	BRE-NW	.6693	.5960	.5018	.6364	.5689	.7778	.8031	.4920	.5258	.6093
	BRE-1T	.5573●	.5440 ●	.3911	.6587	.1387	.2436	.3413	.1067	.1484	.2284
	BRE-MAXT	.5244●	.5084●	.3822	.6489	.1120●	.2311	.3271	.1004	.1227	.1920
60	Mean	.5447	.5613	.4137	.6889	.1859	.3159	.4038	.1277	1922	.2521
	Median	.6439	.6541	.5598	.6302	.5878	.6132	.6244	.5824	.6213	.5711
	AggrList	.5579	.5261	.3778	.6468	.2492	.3373	.4042	.2564	.2780	.3078
	BRE-NW	.5714	.5826	.4820	.6867	.2843	.4082	.5095	.4925	.6270	.6803
	BRE-1T	.5097	.5411	.3955	.6558	.1829	.3151	.3897	.1504○	.2078○	.2734
	BRE-MAXT	.5247	.5362	.3952	.6620	.1797	.3057	.3938	.1439○	.2045○	.2706
120	Mean	.5545	.5359	.4168	.6468	.2202	.3311	.4126	.1487	.2164	.3018
	Median	.6389	.6519	.5191	.6652	.5714	.6246	.6346	.5665	.6365	.5595
	AggrList	.5454	.5162	.3954	.6694	.2698	.3490	.4243	.2742	.2869	.3412
	BRE-NW	.5646	.5566	.4479	.6153	.2755	.4040	.4653	.3043	.3987	.4765
	BRE-1T	.5536	.5180	.3774●	.6319	.1679●	.2694●	.3405●	.1552	.2210	.2887●
	BRE-MAXT	.5592	.5138	.3574●	.6493	.1998	.2967	.3802	.1522	.2162	.2853●
210	Mean	.5434	.5132	.3783	.6558	.1906	.3150	.4179	.1295	.2303	.3158
	Median	.6658	.6683	.5479	.6704	.5988	.6376	.6462	.6050	.6428	.5837
	AggrList	.5488	.5381	.4371	.6577	.3311	.3848	.4639	.3110	.3480	.3960
	BRE-NW	.5546	.5182	.3872	.6420	.2241	.3552	.4535	.2124	.3209	.3998
	BRE-1T	.5744	.4656●	.3292●	.6554	.1679●	.3116●	.3776●	.1167●	.2025●	.2756●
	BRE-MAXT	.5852	.4668●	.3191●	.6649	.1470●	.2428●	.3328●	.1324	.1896●	.2602●

erated Tab. 4.1 shows the $DisJ$ values. The performance has been evaluated using the scaled Spearman footrule distance ($s.F$, Eq. 2.6 in Sec. 2.2) [4] between the $\tau^{T_{rank}}$ and the top-k ranking produced by BRE. We used the same value of k to select the top items on the BRE's output ranking. The result showed in Tab. 4.2 are also plotted in Fig. 4.1, where we compare BRE with respect to each competitors in terms of difference of the $s.F$ distance. In the plots of Fig. 4.2 a negative diffence of the $s.F$ distance means better performance of BRE with respect to the competitors.

We have evaluated the NW and the iterative version of BRE using as true-rank estimator the raw mean. We have evaluated only the raw mean as true-rank estimator, since we have explored the effect of the competitor methods as true-rank estimator in Experiment 1 (Sec. 3.4) on total rankings. The use of the raw mean instead of the mean as true-rank estimator is supported by the same reasons argued for the total rankings, even if we do not show a complete comparison.

Among all the competitors evaluated the mean in Tab. 4.2 shows the best results especially with low k values and $N = 10, 30$. Increasing the number of k, all the competitors decrease their performance in terms of $s.F$ distance with respect to the $\tau^{T_{rank}}$. Also com-



(a) BRE vs. Mean

(b) BRE vs. Median

(c) BRE vs. AggrList

Figure 4.1: Difference of s.F distance of BRE and the competitors. With △ and ∘ are highlighted respectively the cases where BRE outperforms significantly the competitors and BRE is outperformed significantly by the competitors.

paring the performance of the competitors with respect to the *DisJ* coefficient measured in Tab. 4.1, we note that the competitor methods (especially the mean and the *AggrList*), show low $s.F$ values when the sets of items of the rankings are particularly disjointed.

A comparison between BRE-1T and the mean, shows that BRE-1T has some difficulties to estimate ranking especially for low k values. From Fig. 4.1(a) we notice that with $k = 15$ BRE-1T does not outperform significantly the mean in any case with $N = 10$, 30, except for the cases 1 and 2 with $N = 3$. A difficulty of BRE to aggregate top-k rankings of length $k = 15$ is probably derived by the high *DisJ* values showed in this case (see Tab 4.1). On the other hand, increasing the k values ($k = 120, 210$) BRE-1T shows the same outstanding performance against the mean observed in the total ranking experiments. Moreover with with $k = 120, 210$ BRE-1T outperforms the mean also in the case *good* with $N = 10, 30$. This positive trend of the BRE performance is clearly also shown in Fig. 4.1(a).

From Fig. 4.1(b) BRE-1T has significantly better performance with respec to the median in all the cases (good, equal, poor) and for all the k values. BRE-1T outperforms significantly the *AggrList* in all the cases with $N = 30$ across all the k values (Fig.4.1(c)). With respect to $N = 10$ BRE-1T outperforms significantly the *AggrList* in all the three cases but only for $k = 120, 210$.

From Tab. 4.2, we point out that BRE-MAXT shows quite good result with respect to the BRE-1T version. For $k = 15$ in the case *good*, BRE-MAXT show an impressive improvement of the scaled distance (.1120) with respect to BRE-1T (.1387). The iterative schema seems particularly effective also when k is high, since for $k = 210$ BRE-MAXT outperforms the BRE-1T version in the cases considered ($N = 10, 30$). A possible explanation of this effect is that the replacement of the top-k rankings with a total ranking (in U^* set) transforms progressively the combination of top-k into a combination of total rankings.

The not weighted version of BRE, has showed very poor results for all the k values compared with the iterative version (BRE-1T). BRE-NW has globally worst performance with respect to BRE-1T in all the cases and for all the N values. With $N = 10, 30$ in all the cases, the difference of the distance obtained by BRE-NW and BRE-1T decreases as the k values increase. This lead us to suppose that the weight schema gives a positive contribution also in the case with high $DisJ$ values and low k values.

Although BRE does not provide a constant improvement of performance with respect to the mean with low k values, we conclude that BRE exhibits competitive performance on the aggregation of top-k rankings with respect to the median and the *AggrList*. The limited performance of BRE in case of very short lists can be caused by the weight schema used, the effect of the distance for the computation of the weights will be evaluated in the next experiments. We recall that in this experiment we have not evaluated the role of the true-rank estimators in the top-k rankings scenario, a detailed investigation will be considered in future work.

4.4 Experiment 2: BRE on Partial Rankings

This experiment addresses the estimation of the true ranking in the case of partial rankings, when the experts have not a complete knowledge of the universe set U. For this empirical scenario, we have generated synthetic data that meet the hypothesis of partial ranking discussed above (Sec. 4.1). More formally, we have N partial rankings τ^j provided by experts where $\forall j \in 1, \ldots, N\ S_j \subset U$, $|\tau^j| = |S_j|$. Each experts has as input S_j a subset of the U set, and $\forall j \in 1, \ldots, N\ |S_j|$ is fixed , so the corresponding rankings

are simply total rankings on each set S_j with $|\tau^j| = |S_j|$.
As in Experiment 1 (Sec. 4.3), the competitor methods are the mean, the median and the approximate solution of the Footrule optimal aggregation based on the cross-entropy approach (*AggrList*). Again, we investigate the iterative version of BRE (BRE-MAXT), the weighted version (BRE-1T) and the not weighted one (BRE-NW). As in the previous experiment (Sec. 4.3) the true-rank estimator is the raw mean.
The generation of the data for all the quality cases and for all N values ($N = 3, 10, 30$) starts from a fixed true ranking $\tau^{T_{rank}}$ of 300 items ($n = 300$) as follows:

- For each N value, we have generated the base ranking of each expert, as a random ranking on a random subset of items of U. The length of each base ranking is equal to 120.
- From the N base rankings, we generate the input rankings as random permuted rankings according to the values of Spearman coefficient ρ relative to the quality cases evaluated (*case 1,case 2, case 3, case 3, good, poor, equal*, see Sec. 3.4). The input rankings generated τ^j have length of 120, and are total rankings with respect to the correspondent base rankings.

For each case and for each N we have generated 10 independent replicas using the same parameters in order to assess the statistical significance of the results. Like in previous experiments, we have used a paired two-tailed t-Test with $\alpha = 0.05$. The performance are evaluated with the scaled Spearman footrule distance ($s.F$) from $\tau^{T_{rank}}$. The length of the partial lists could be also investigated, but we have decided to set the length of the partial rankings ($k = 120$) since the estimation of the true ranking on this hypothesis is quite difficult. We remark that the base rankings are randomly generated from $\tau^{T_{rank}}$, but the input rankings are still permutations from the base rankings, so the relation between the generated input rankings and the true ranking is more complex than in the total-ranking cases. Like in the previous experiment (Sec. 4.3), on the result showed in Tab. 4.3 we have also plotted in Fig. 4.2 the differences of BRE with repesct to the competitors in terms s.F. The significance of the result is shown in Fig. 4.2 and also included in the discussion.

The difficulty of the task generated for this partial rankings setting is showed by the results in Tab. 4.3. The three competitors obtain on average very high distance from the true ranking independently of the quality cases and the number of rankings considered. With $N = 30$ and $N = 10$ for low values of $DisJ$ coefficients ($DisJ$ = .2458, $DisJ$ = .0830) the distance obtained by the competitors is higher than the distance obtained in the Experiment 1 (Sec 4.3) in the same conditions of $DisJ$ and of N values (see Tab. 4.2 with k=120). From that premise, we expect that also BRE does not provide excellent results since its performance are strongly related to the correlation of the true-rank estimator with the true ranking.
In all the four cases with $N = 3$, BRE-1T outperforms significantly the *AggList* method in case 1, 2, 3, but with respect to the other two competitors BRE does not show significant difference in terms of $s.F$ distance (Fig. 4.2). Increasing the number of rankings to $N = 10$, BRE-1T outperforms only the *AggList* in the case *good* whereas the other competitors do not have any significant improvement of performance except in the case

Table 4.3: Partial rankings: Average of the scaled Spearman footrule distance ($s.F$) of BRE and of the competitor methods with respect to the true ranking. For each case the average of the $DisJ$ coefficent is showed.

	Evaluation measure: $s.F$									
	N=3				N=10			N=30		
Method	1	2	3	4	good	equal	poor	good	equal	poor
	min DisJ=.30				min DisJ=.10			min DisJ=.03		
$DisJ$	.6472				.2458			.083		
Mean	.6784	.6934	.7022	.6802	.6763	.7046	.6959	.6577	.6423	.6612
Median	.6782	.6441	.6963	.6545	.6666	.6586	.6567	.6517	.6337	.6743
Aggrlist	.7167	.7259	.7180	.7242	.7036	.6898	.6964	.6351	.6467	.6472
BRE-NW	.6758	.6610	.6980	.6510	.6749	.6995	.6807	.6575	.6571	.6577
BRE-1T	.6837	.6672	.7066	.6803	.6724	.7032	.6784	.6717	.6474	.6571
BRE-MAXT	.6683	.6708	.6951	.6829	.6762	.6772	.6671	.6791	.6593	.6690

equal when the median obtains the best result (.6586) . As an overal considerations from Fig. 4.2, we notice that the median shows the best performance with respect to BRE, and *AggList* is the competitors on which BRE show the better results. With $N = 30$ the *AggrList* and the mean outperform significantly BRE-1T in the case *good*.
The performance of iterative version is really similar to the weighted version ($BRE-1T$), except for the case *poor* with N=10 where BRE-MAXT outperforms the mean and the *AggrList* method. In Experiment 1 (Sec 4.3) BRE-NW has showed an important gap of performance with respect to the weighted one but in this experiment the gap of performance is strongly reduced also in the case *poor* where the weighting schema has always played an important role.

Although the BRE does not outperform all the competitor methods in this partial ranking setting, we have confirmed that BRE is strongly based on the quality of the true-rank estimator. In the case of total rankings the ranking produced by the true-rank estimator method is always well related to the true ranking, but in the partial rankings we have that the rankings produced by the true-rank estimator methods on the items U^* cannot have any particular relation with the true ranking. This issue probably gets BRE to produce rankings that are not so good estimations of the true ranking. As just discussed in the previous chapter, the use of *a priori* knowledge about some items could be a valid support to integrate the true-rank estimator method for the quality estimation of the input rankings. As future work, the role of true-rank estimator in partial rankings will be investigated.

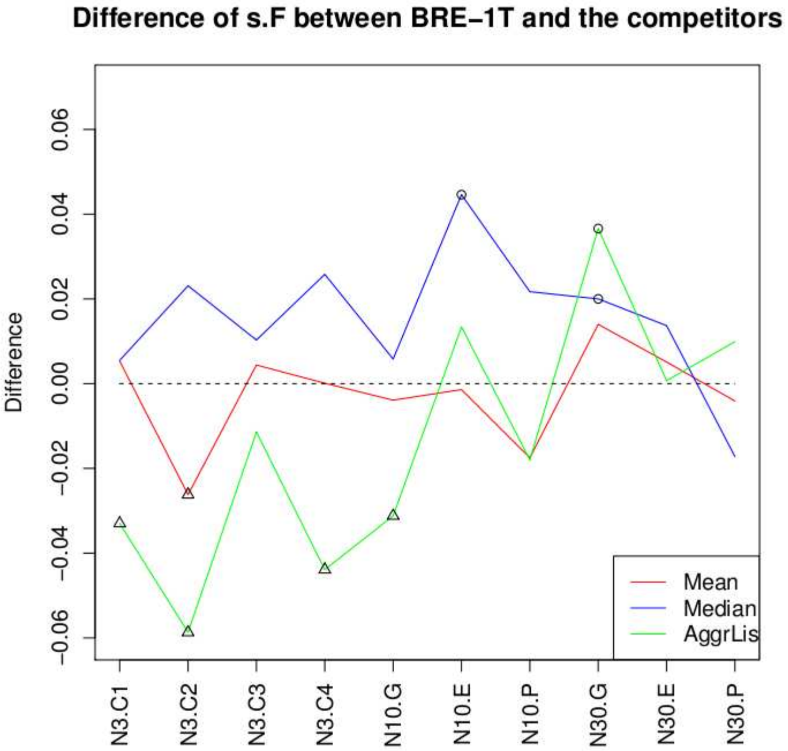


Figure 4.2: BRE on Partial Rankings: Differences of BRE with respect to competitors in terms of s.F distance. With △ and ∘ are highlighted respectively the cases where BRE outperforms significantly the competitors and BRE is outperformed significantly by the competitors.

4.5 Experiment 3: Top-k Lists of Partial Rankings

In the previous experiments we have evaluated BRE in two different settings: the top-k rankings where the experts share the entire universe and the partial rankings in which the experts know only small subsets of U. The possible situations of rankings (partial/total) with respect to their S_j set are too many to be extensively evaluated in this work. Before moving on real data, we have decided to evaluate BRE in a situation that is half way between the two previous experiments. We have N top-k rankings τ^j provided by experts where $\forall j \in 1, \ldots, N \, S_j \subset U \, l_j = |S_j|, k_j = |\tau^j|, k = k_j$. Each expert has as input S_j a subset of items of the U set, and $|S_j|$ is still fixed but the corresponding rankings are top-k rankings with length k where $\forall j \in 1, \ldots, N \, k = k_j$. The main difference with the setting in Experiment 2 (Sec. 4.4) is that the output rankings are top-k rankings with respect to the set S_j. Moreover, the task of the estimation of the true ranking is more difficult of the task in Experiment 2 (Sec. 4.4), due to the top-k rankings from the generated rankings.

The rankings are based on the same data generated in Experiment 2 (Sec. 4.4). The only difference is that we have selected the first 60 items from the perturbed rankings generated by the base rankings of length 120. The competitors and the versions of BRE

Table 4.4: Top-k rankings from partial lists: Averaqe $s.F$ distance of BRE and of the competitor methods with respect to the true ranking. For each case the average of the $DisJ$ coefficent is showed.

	Evaluation measure: $s.F$									
	N=3				N=10			N=30		
Method	1	2	3	4	good	equal	poor	good	equal	poor
	min DisJ=.30				minDisJ=.10			min DisJ=.03		
DisJ	.8044	.815	.8211	.8311	.4460	.4457	.4458	.16	.16	.16
Mean	.6895	.7084	.7002	.6911	.6700	.6754	.6627	.6370	.6782	.6089
Median	.6479	.6671	.6822	.6438	.6708	.7004	6456	.6692	.6826	.6863
AggrList	.6726	.7018	.6927	.6759	.6903	.6653	.6698	.6613	.6827	.6430
BRE-NW	.6773	.7154	.6994	.6766	.6420	.6728	.6844	.6344	.6354	.6260
BRE-1T	.6738	.7044	.7157	.6900	.6529	.6679	.6671	.6604	.6461	.6053
BRE-MAXT	.6953	.6843	.6983	.6844	.6426	.6553	.6706	.6582	.6481	.5974

(NW, 1T, MAXT) are the same of Experiment 2 (Sec. 4.4). The results in Tab. 4.4 show the scaled Spearman footrule distance of the output rankings from the true ranking and the $DisJ$ coefficients for each generated case. In Fig. 4.3 are plotted the difference of BRE with respect to competitors in terms of $s.F$ distance. The statistical significance of the results obtained is show in Fig. 4.3 and included in the discussion. As in the previous experiments, we have used a paired two-tailed T-test with $\alpha = .05$ to asses the significance of the results in various cases.

The first comment regards the increasing of the $DisJ$ coefficient in all the cases (especially with $N = 10, 30$) with respect to Tab. 4.2. As our expectation, this a direct effect of the increased partiality of the input rankings. This effect increases the difficulties of BRE to provide good results, as also mentioned in the Experiment 1 (Sec. 4.3) (for $k = 15, 60$). The performance of BRE-1T are significantly better than the median in case *equal* with $N = 10$ and in the case *poor* with $N = 30$. From Fig. 4.3, BRE-1T shows superior performance than AggrList in the case *good* ($N = 10$) and in the case *equal* (N=30). The mean is outperformed significantly only by the iterative version in the case *equal* with $N = 30$. In all the other cases the difference of distance of BRE and the competitors are not significant.

From this result we conclude that BRE is really sensitive to the input rankings that have very few elements in common (high $DisJ$ values). However, in this difficult task BRE-1T and BRE-MAXT have showed interesting performance with respect to the competitor methods.

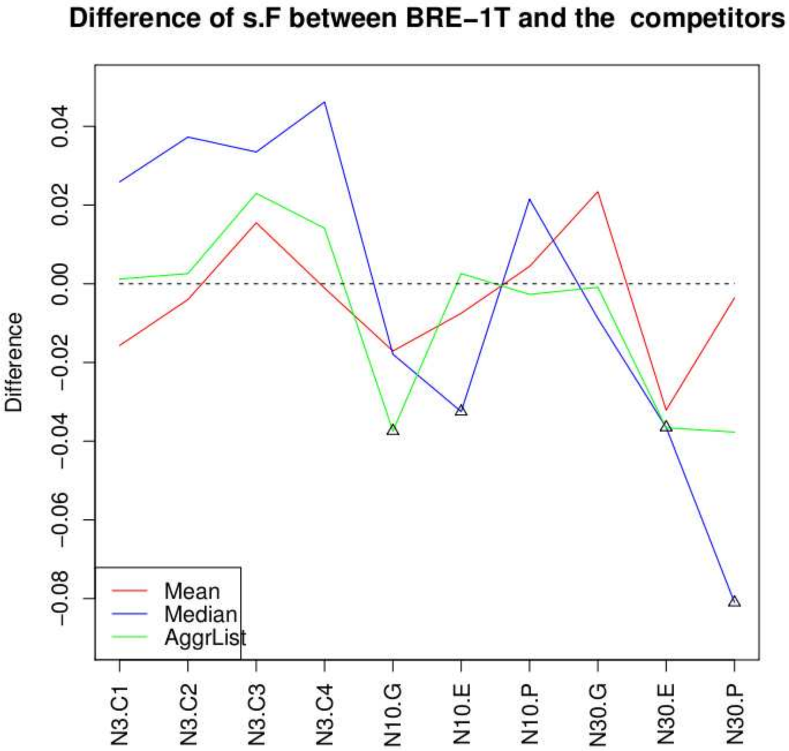


Figure 4.3: BRE on Top-k Lists of Partial Rankings: Differences of BRE with respect to competitors in terms of s.F distance. With △ and ∘ are highlighted respectively the cases where BRE outperforms significantly the competitors and BRE is outperformed significantly by the competitors.

4.6 Synthethic Data: Conclusions

From the definition of the partial/top-k rankings described in this chapter, we have modeled through synthetic data three possible settings on which we have evaluated BRE, and the other competitors on the task of the estimation of the true ranking. In all the experiments, we have used a straightforward version of BRE, based on the raw mean as true-rank estimator, and Spearman footrule distance to the weights computation. We have decided to not evaluate the different distances for the weight and other bba assignment, since we were more interested to explore how the problem of the estimation of the true ranking on partial rankings could be really complex for the BRE algorithm due to the lack of a priori knowledge about the experts involved.
The connection of the rankings in input with the true ranking underlyng the problem, is the key point for the performance of BRE. The lack of this relation brings the true-rank estimator to be not enough accurate for the quality estimation of the input ranking. In Experiments 2 (Sec. 4.4) and 3 (Sec. 4.5), the performance of BRE are more or less the same for all the quality cases considered. The role of the true estimator methods with respect to the true ranking and its possible impact on BRE performance is quite complex to investigate in this work due to huge number of partial rankings that can be generated.

The limitation of the true-rank estimator with respect to the quality of rankings will be investigated on real data, using different distance functions for the weights computation.

The length of the partial rankings is another important issue, since short input rankings have few elements in common. This increase the possibilities to have a lot of ties in the output rankings, since the items not common in the input rankings will have similar belief functions on Θ and the combination does not change their belief. In this case only the discount step updates the beliefs of the items that are not common among the rankings.

4.7 LETOR Benchmark

LETOR [3] is a collection of datasets for benchmarking algorithms related to the learning to rank problem [26]. From the latest version, LETOR 4.0 includes also the setting for the ranking aggregation problem. LETOR contains two query sets from Million Query track of TREC 2007 and TREC 2008 that count respectively 1800 and 800 queries with annotated documents. The task is to combine rankings of documents from search engines retrieved on queries, in order to obtain a ranking that has the most relevant documents in the higher positions.

We have decided to evaluate the BRE algorithm on LETOR, since it provides the baseline results of some state-of-the-art methods and a solid evaluation tool that avoid evaluation problems of the results. For the task of rankings aggregation, LETOR provides the results of Borda Count and a method based on a probabilistic model on permutations called CPS [19]. Since the latter method is based on the training step to generate a probabilistic model, we focus on Borda Count method that is an heuristic totally fair as competitor for the BRE algorithm. Since the Borda Count methods can include several aggregation functions, we have applied the Borda Count (mean of the rankings) like in previous experiments and we have found the same results of the Borda Count showed in LETOR. As a consequence, the mean has been used as the true-rank estimator method inside BRE. Although, we have used only one competitor in this experiments, the mean (Borda Count) shows good results in LETOR so we expect that it can be a valid and informative true-rank estimator.

4.7.1 LETOR Dataset and Evaluation Measures

LETOR dataset is composed of two sets of queries 2007-agg and 2008-agg respectively composed of 21 and 28 input lists. For each query the task is to aggregate the input top-k lists. The number of documents of each query ranges from 8 up to 40. In each query, the input lists are top-k rankings with different values of k. We point out that this is a remarkable difference from the previous experiments (Sec. 4.3, 4.4, 4.5) where the length of the input rankings has been fixed for all the rankings. Each query corresponds to an independent aggregation task, even if the performance are computed as average among all the queries.

A query Q is composed of N top-k rankings τ^j of length k_j, and the number of all documents ranked in the N rankings is denoted as Dn. As in the previous experiments we have used the $DisJ$ coefficient (Eq. 4.1) to measure the common items present in the input rankings, for LETOR we introduce the Partiality Index (P.I) to measure the degree of partiality of the rankings. Given a query Q the $P.I$ is defined as:

$$P.I(Q) = \frac{\sum_{j}^{N} k_j}{Dn * N}$$

where k_j is the number of items contained in the j-th ranking. The value of the $P.I$ are bounded by $\frac{1}{N} \leq P.I(Q) \leq 1$, where $P.I(Q) = \frac{1}{N}$ means that all the lists rank all the documents of the query and $P.I(Q) = 1$ is the case when all rankings have $k_j = 1$ (extreme

Table 4.5: Average Partiality Index (*P.I*) for the dataset 2007-agg and 2008-agg

	2007-agg	2008-agg
Fold1	.6624	.4137
Fold2	.6578	.3990
Fold3	.6631	.4050
Fold4	.6528	.4066
Fold5	.6555	.4067
Average	*.6583*	*.4060*

case). The Partiality Index (*P.I*) gives an idea of how the rankings of the queries are partial with respect to the number of documents present in the queries. We refer to this concept as the partiality of the input rankings. From Tab. 4.5 we notice that the queries in the dataset 2008-agg show an high partiality with respect to the dataset 2007-agg. We expect that the BRE will meet more difficulties on 2008-agg than on 2007-agg.

The evaluation of the performance on LETOR is based on *precision* [26], *mean average precision*[26] and *normalized discounted cumulative gain* (NDCG)[50]. All these well-known measures are used to evaluate performance in information retrieval . There are three levels of relevance for the documents in both the datasets: highly relevant, relevant, irrelevant.

Precision Given a list of documents for a query, precision at n is defined as:

$$P@n = \frac{\# \, relevant \, docs \, top \, n \, positions}{n}$$

with the precision is evaluated only a binary judgment, relevant or not relevant, in top n documents provided by the list. The precision does not make distinction between the highly relevant and relevant documents.

Mean Average Precision For a query, the average precision for all the documents Dn, is defined as:

$$AP = \frac{\sum_{i=1}^{Dn} P@i * rel(i)}{\# \, total \, relevant \, document \, for \, this \, query}$$

$$rel(i) = \begin{cases} 1 & \text{if } i\text{-th doc is relevant} \\ 0 & \text{otherwise} \end{cases}$$

where $rel(i)$ is function on the relevance of i-th retrieved documented.

NDCG: Normalized Discounted Cumulative Gain The NDCG is a measure of ranking quality, that takes into account the position and the relevance of documents in the provided list [50]. Moreover, it can handle multiple relevance values instead of just the binary case. NDCG is a sort of weighted precision based on logarithmic scale, that penalizes more the change of position of the highly relevant documents versus non relevant documents. The NDCG at the position n is computed [50]-[51] as following:

$$NDGC@n = Z_n \sum_{i}^{n} \begin{cases} 2^{r(i)} - 1 & j = 1 \\ \frac{2^{r(i)}-1}{\log(i)} & \text{j} > 1 \end{cases}$$

where $r(j)$ is the relevance of i-th document, and Z_n is a numerical constant in order to have for the perfect list $NDCG = 1$. Also the mean of $NDGC@n$ with respect to all the n documents retrieved by a single query is computed.

For each evaluation measure $P@n$ MAP, $NDCG@n$, the results are computed as average for all the queries presented in the dataset. The larger the NDCG value and precision value, the better the aggregation accuracy.

The performance are computed by comparing the order of the documents in the output rankings with respect to the level of relevance of the document. The goal is to produce a ranking that contains the more relevant documents in the higher positions. We point out that LETOR does not really have an underlying true ranking, in fact the order of the documents with the same level of relevance does not matter in terms of evaluated performance.

4.7.2 The BBA and the Weights Computation Evaluated

With respect to the LETOR dataset, we will evaluate some modifications of BRE in order to encode different knowledge to tackle the task of the LETOR dataset. These modifications regard the belief basic assignment and the distances used to compute the weights.

Let we denote as U_q the set of documents present in a query where its cardinality is $Dn = |U_q|$, and τ^* and τ^j are respectively the augmented and the partial rankings in input. Finally, the length of the partial rankings is $k_j = |\tau^j|$. For LETOR, the augmented rankings are set as follows:

$$\tau^{*j}(i) \begin{cases} \tau^j(i) & if \quad x_i \in C_j \\ \frac{\sum\limits_{r=k_j+1}^{Dn} r}{Dn-k_j} & if \quad x_i \in U_q/C_j \end{cases} \tag{4.6}$$

where C_j is the set of ranked items in τ^j. The rank associated with the document $x_i \in U_q/C_j$ is the mean of the $Dn - k_j$ missing-rank values and $k < \frac{\sum_{r=k_j+1}^{Dn} r}{Dn-k_j} < Dn$. We recall the bba assignment used for partial rankings in the previous experiments:

$$\begin{array}{ll} \text{if} \quad x_i \in C_j & \text{if} \quad x_i \in U_q/C_j \\ m_{ji}(P) = \dfrac{M1 - (\tau^{*j}(i) - 1)}{M2} & m_{ji}(P) = 0 \\ m_{ji}(\neg P) = 0 & m_{ji}(\neg P) = 0 \\ m_{ji}(\Theta) = 1 - \dfrac{M2 - (\tau^{*j}(i) - 1)}{M2} & m_{ji}(\Theta) = 1 \end{array} \tag{4.7}$$

In order to give a clearly presentation of the different belief assignments evaluated we have introduced two variables $M1$ and $M2$ in Eq. 4.7 that take different values in each belief assignment as described in the following:

Base $M1$ and $M2$ are equal to k_j. For the k_j elements that belongs to τ^j the belief on P decreases linearly with the length k_j. This is the same bba used in previous experiments (Sec. 4.3,4.4,4.5) and also for the total rakings (Eq. 3.1).

M $M1$ and $M2$ are set to $\frac{\sum_{r=k_j+1}^{Dn} r}{Dn-k_j}$. With respect to the assignment Base, the mass on P is smoother for the last k items, due to the fact that $k < \frac{\sum_{r=k_j+1}^{Dn} r}{Dn-k_j} < Dn$.

Opt 1 $M1$ is equal to k_j and $M2 = Dn$. This assignment models the assumption that rankings with a small k_j with respect to Dn, are not so reliable as rankings with a number of elements k near to Dn.

Opt 2 $M1$ is equal to $\frac{\sum_{r=k_j+1}^{Dn} r}{Dn-k_j}$ and $M2 = Dn$. This provides a smoother effect than Opt1. With this bba, also rankings with low k_j values are mildly penalized with respect to rankings with an high number of items.

For all the four bbas evaluated, we also introduce the following modifications. For all the first items (rank value equals 1) that received a $m(P) = 1$, we slightly reduce the $m(P)$ by an ϵ value. Our expectation for this modification is that the risk of ties in the final output is reduced. In this case we use the same value ($\epsilon = 0.005$) for all the rankings, but different ϵ values could be used for the each input rankings if external information about the experts are available.
The effect of the four bbas previously described are showed in Fig. 4.4, where they are applied on a ranking with $k_j = 4$ of a query with $Dn = 8$. We remark how the belief on P for the the first item, is drastically decreased from the bba *Base* and K with respect to the bba *Opt 1* and *Opt 2*.

The different ways to compute the weights introduced in BRE are the following:

i.F The weight of each ranking is computed using the induced Spearman footruke distance.

$$w_j = i.F(\tau^{TE}, \tau^j)$$

The induced Spearman footrule distance, described in Eq. 2.5, is a distance between a partial and a total ranking. The partial ranking are the input rankings (τ^j) and the mean of the rankings is the total ranking over Dn documents present in the query (τ^{TE}).

s.F The weight of each ranking is computed using the scaled Spearman footruke distance.

$$w_j = s.F(\tau^{TE}, \tau^j)$$

Same consideration also for the scaled Spearman footrule (Eq. 2.6) between a input partial ranking (τ^j) and the mean of the rankings (τ^{TE}).

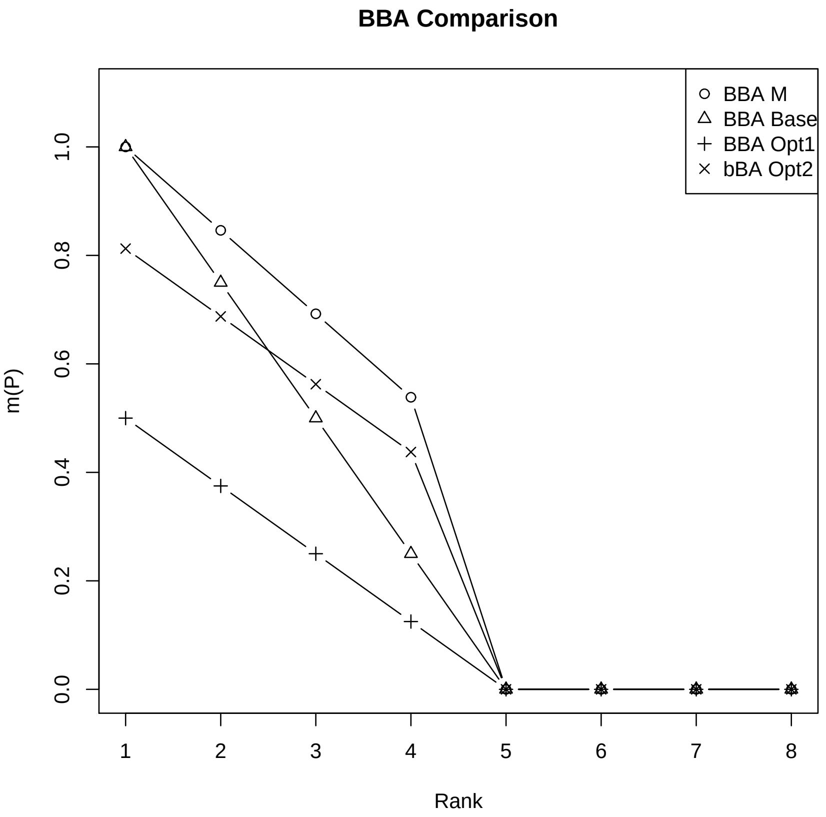


Figure 4.4: The four bbas evaluated on LETOR, applied on a ranking with $k_j = 4$ where the query contains $Dn = 8$ documents

FSum The weight for each τ^{*j} is computed as follows:

$$w_j = \frac{\sum_{l=1}^{10} D^l(\tau^{*j}, \tau^{TE})}{10}$$

where D^l is defined as

$$D^l = F(\tau^{*j}_{|\{Top\, l\, x_i \in \tau^{TE}\}}, \tau^{TE}_{|\{Top\, l\, x_i \in \tau^{TE}\}})$$

In D^l the F distance is computed between the input ranking τ^{*j} selecting the first top l items of τ^{TE} and the top l items contained in the ranking provided by the true-rank estimator method (τ^{TE}). The idea is to compute the weight w_j adding progressively only the top l items present in the true-rank estimator (the mean), since we expect that in the top-l elements of the mean are contained the most relevant items. The number of top items l is fixed to 10, since we are looking to increase the performance of P@n and NDSG@n for the first items retrieved.

Table 4.6: LETOR 2007-agg: Precision and NDCG results of the BRE NW for all the bbas evaluated against the mean.

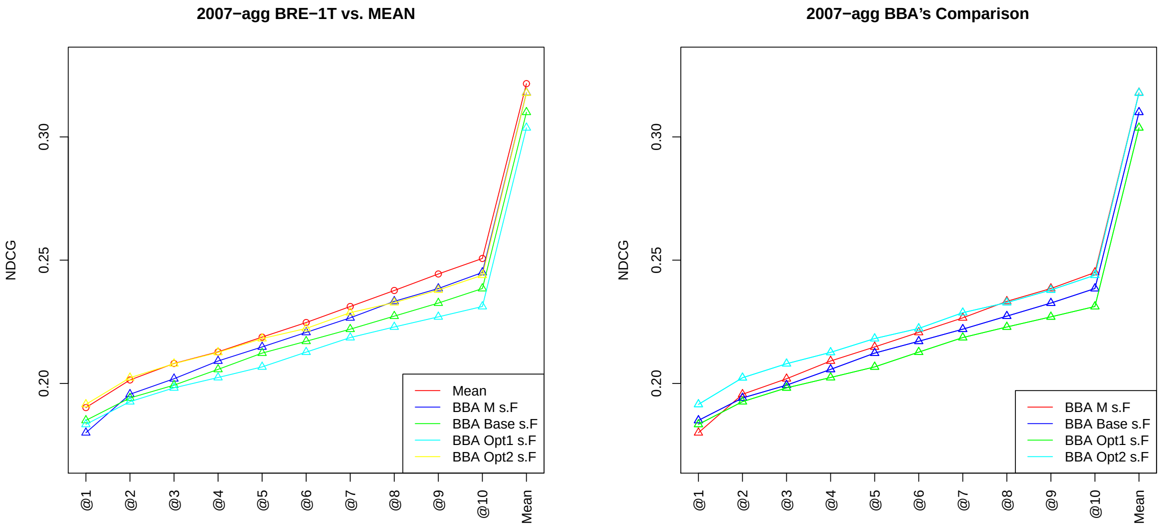
Meth.	BBA W.	P@1	P@2	P@3	P@4	P@5	P@6	P@7	P@8	P@9	P@10	MAP
Mean		.2488	.2524	.2569	.2580	.2597	.2626	.2645	.2666	.2684	.2690	.3252
NW	M	.2210	.2258	.2215	.2218	.2236	.2266	.2298	.2335	.2370	.2382	.3063
NW	Base	.2169	.2140	.2173	.2177	.2191	.2230	.2237	.2281	.2315	.2333	.2994
NW	Opt1	.2140	.2234	.2270	.2292	.2311	.2326	.2360	.2371	.2378	.2388	.3020
NW	Opt2	.2193	.2276	.2286	.2314	.2340	.2378	.2393	.2430	.2439	.2473	.3100
							NDCG					
		@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	Mean
Mean		.1902	.2014	.2081	.2128	.2188	.2247	.2312	.2377	.2444	.2507	.3216
NW	M	.1710	.1802	.1825	.1866	.1911	.1968	.2025	.2098	.2170	.2227	.3018
NW	Base	.1649	.1692	.1755	.1790	.1842	.1901	.1946	.2020	.2087	.2149	.2918
NW	Opt1	.1616	.1754	.1815	.1879	.1932	.1976	.2027	.2074	.2123	.2173	.2914
NW	Opt2	.1692	.1802	.1850	.1909	.1969	.2033	.2087	.2149	.2207	.2282	.3022

Using the $i.F$ and $s.F$ distances to compute the weights, the ranking of the mean is projected with respect to the set of elements of the partial rankings (τ^j). For the weights computed by $FSum$, the comparison works in the opposite way the input rankings are projected with respect to the top 10 items of the mean.

4.7.3 LETOR: Results and Discussion

We have split the results of each dataset in two tables, the first contains BRE with the not weighted version (Tab. 4.6 and Tab. 4.7), and in the second contains the results of BRE-1T and BRE-MAXT (Tab. 4.8 and Tab. 4.9). For the most intersting results of BRE with the bba's and the distances evaluated we have plotted the graphs showed in Fig. 4.5 and Fig. 4.6.

With respect to the 2007-agg dataset, The results of BRE-NW are very interesting in order to understand the effect of bba on BRE. From Tab. 4.6, BRE-NW results are far from the mean in terms of precision and NDCG, but we can notice that the bba Opt2 works better than the Opt1, and the bba M better than the Base one. Taking into account the BRE-1T result in Fig. 4.5(b), we notice that the most effective bba's are the Opt2 and the M one. From the comparison of BRE-1T with respect to the mean shown in Fig. 4.5(a) we notice that BRE-1T does not outperform globally the mean, but we highlight that the bba Opt2 gives a slightly increment of the NDCG values for $n < 3$. The comparison of the different distance evaluated is showed in Tab 4.8. Taking into account the bba M and the bba Opt1, the increasing of performance among the distances evaluated is really interesting, and it shows how the weights are one of the key points of BRE. The F distance does not work well as the $i.F$, $s.F$ and $Fsum$, in particularly BRE with the F distance has low precision at small n. Instead the $Fsum$ shows very good performance with the bba M and bba Opt2.We include the results of the iterative version only for two configurations (bba M and Opt2 with $Fsum$ distance), that show how BRE-MAXT slightly improves the 1T version only with Opt2 assignment. Although we have not showed the results of BRE-MAXT for all the configurations the iterative version does not improve the 1T version enough to outperform the mean.



(a) BRE vs. Mean

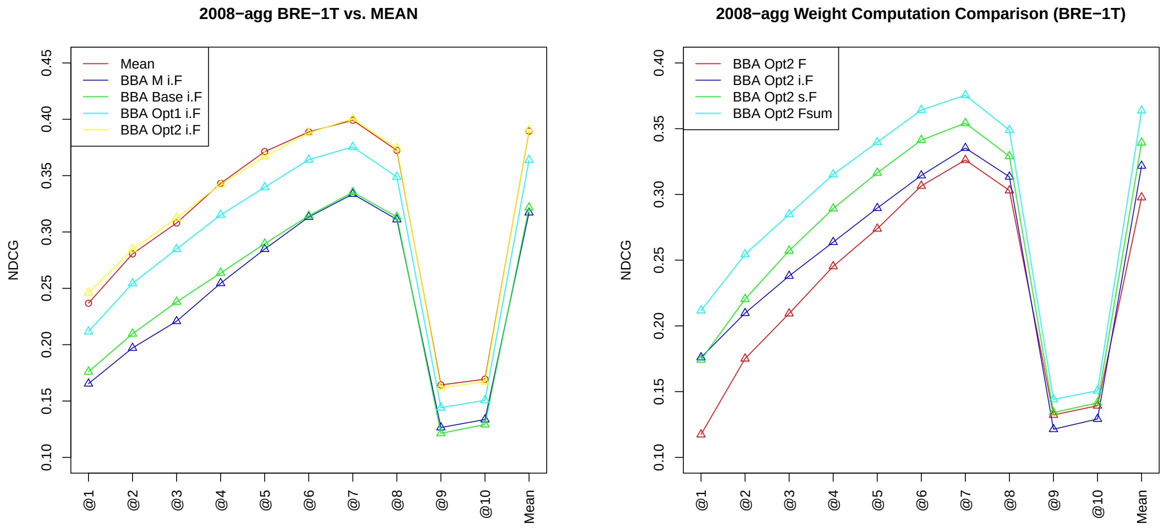
(b) BBA's Comparison

Figure 4.5: 2007-agg: Result of BRE among the different BBA's evaluated in terms of NDCG

Table 4.7: LETOR 2008-agg: Precision and NDCG results of the BRE-NW for all the bba evaluated against the mean

Meth.	BBA W.	P@1	P@2	P@3	P@4	P@5	P@6	P@7	P@8	P@9	P@10	MAP
Mean		.2972	.3042	.2938	.2975	.2903	.2783	.2642	.2503	.2367	.2230	.3945
NW	M	.1760	.1977	.2049	.2098	.2107	.2103	.2118	.2062	.1943	.1851	.3042
NW	Base	.2398	.2366	.2432	.2373	.2316	.2292	.2241	.2163	.2056	.1953	.3379
NW	Opt1	.2207	.2411	.2488	.2523	.2492	.2464	.2398	.2290	.2178	.2078	.3448
NW	Opt2	.2309	.2672	.2713	.2771	.2681	.2621	.2516	.2396	.2265	.2152	.3672
							NDCG					
		@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	Mean
Mean		.2368	.2806	.3080	.3432	.3713	.3888	.3992	.3724	.1643	.1694	.3895
NW	M	.1437	.1816	.2131	.2411	.2664	.2873	.3097	.2882	.1088	.1148	.2938
NW	Base	.1896	.2234	.2576	.2790	.3003	.3217	.3393	.3169	.1210	.1274	.3283
NW	Opt1	.1646	.2168	.2493	.2810	.3078	.3325	.3477	.3224	.1320	.1386	.3316
NW	Opt2	.1790	.2509	.2826	.3177	.3426	.3646	.3780	.3530	.1514	.1583	.3612

With regard to the 2008-agg dataset, the situation is the same and the mean is not outperformed by BRE, but interesting considerations arise from the fact that the different bba and distances deal with the low $P.I$ index present in the dataset (Tab. 4.5). From Tab. 4.7, we notice that the bba Opt2 is the best assignment with respect to the other, and also to the bba M. Increasing the number of elements computed by the NDCG and the precision measures, the BRE-NW with bba Opt2 constantly decreases the gap with the mean but not enough to outperform it (Tab. 4.9). As shown in Fig. 4.6(a), BRE-1T with the bba Opt2 shows clearly the best performance in terms of NDCG with respect to the other bba's evaluated. We point out that BRE-1T with Opt2 and the induce footrule as distance outperforms the mean with low margin in NDCG@1, @2, @3, @7, NDCG@Mean



(a) BRE vs. Mean

(b) Weight Computation Comparison

Figure 4.6: 2008-agg: Result of BRE among the different BBA's evaluated in terms of NDCG

(Fig. 4.6(a)). From a comparison of different distances evaluated showed in Fig. 4.6(b), we highlight also the limits of F, instead of the $s.F$, $i.F$ and $Fsum$ that capture better the quality of the input rankings and increase the performance.
As for the 2007-agg dataset, in Tab. 4.9 are included the MAXT result for all the most interesting configuration evaluated. BRE MAXT with both the best bbas shows little improvements with respect to the 1T version when the performance are evaluated on the first 10 elements, instead shows a valuable improvement when the entire ranking produced is evaluated (NDCG@Mean, MAP).

In this experiment BRE has not shown impressive performance with respect to the mean, however we have showed the role of bba in BRE in a real case. The extreme partiality of the rankings in the 2008-agg, has been well modeled in the bba Opt2 that takes into account the length of the rankings with respect to the number of documents. Moreover also the distance used to compute the weights play a relevant role in BRE when partial rankings are involved. To analyze the reasons of the flaw performance of BRE where the mean is used as true-rank estimator we have to consider the following points:

1. Each query is independent, but the performance is computed as the average on all the queries. Since BRE performance are not so bad with respect to the competitor, we believe that there are some queries in which the mean does not provide enough information as true-rank estimator, and also cases where BRE does not improve the mean for other reasons that should be investigated.

2. The partial aspect of the true ranking searched for this task, introduces a notable effort for our algorithms. Empirical evidence of this fact is that BRE obtains interesting performance when the entire length of output rankings is evaluated (MAP and

NDGC@Mean) instead of low performance when only the first n items are evaluated. This issue is probably still related to the problem of global weights vs. local weight (Sec. 3.11).

In the next experiment to empirically verify our consideration discussed at point (1) we have to analyze how this two datasets are composed in terms of which queries are well aggregated by the mean. We will cluster the queries in the dataset with respect to the mean performance, evaluating how BRE works in the various quality groups.

Table 4.8: LETOR 2007-agg: Precision and NDCG results of the BRE for all the bba and the distances evaluated against the mean

Meth	BBA	W.	P@1	P@2	P@3	P@4	P@5	P@6	P@7	P@8	P@9	P@10	MAP
Mean			.2488	.2524	.2569	.2580	.2597	.2626	.2645	.2666	.2684	.2690	.3252
1T	M	F	.2028	.2211	.2308	.2370	.2398	.2426	.2469	.2491	.2515	.2526	.3146
1T	M	i.F	.2334	.2438	.2471	.2522	.2551	.2568	.2590	.2618	.2616	.2636	.3249
1T	M	s.F	.2281	.2426	.2471	.2506	.2535	.2554	.2583	.2603	.2599	.2613	.3211
1T	M	Fsum	.2388	.2421	.2497	.2518	.2584	.2617	.2634	.2641	.2648	.2652	.3247
MAXT	M	Fsum	2376	.2536	.2507	.2529	.2567	.2592	.2597	.2593	.2621	.2620	.3225
1T	Base	F	.2369	.2438	.2443	.2474	.2520	.2511	.2531	.2540	.2541	.2554	.3159
1T	Base	i.F	.2405	.2470	.2475	.2525	.2553	.2571	.2595	.2621	.2618	.2637	.3248
1T	Base	s.F	.2369	.2438	.2443	.2474	.2520	.2511	.2531	.2540	.2541	.2554	.3159
1T	Opt1	F	.2376	.2341	.2374	.2391	.2374	.2398	.2435	.2468	.2482	.2490	.3098
1T	Opt1	i.F	.2382	.2447	.2430	.2431	.2439	.2466	.2482	.2506	.2519	.2525	.3127
1T	Opt1	s.F	.2405	.2447	.2461	.2456	.2447	.2479	.2511	.2510	.2503	.2497	.3125
1T	Opt2	i.F	.2447	.2492	.2512	.2527	.2576	.2590	.2604	.2612	.2620	.2635	.3229
1T	Opt2	s.F	.2483	.2545	.2563	.2565	.2569	.2565	.2589	.2580	.2586	.2599	.3224
1T	Opt2	Fsum	.2477	.2480	.2493	.2535	.2580	.2591	.2608	.2614	.2627	.2635	.3217
MAXT	Opt2	Fsum	.2524	.2501	.2530	.2567	.2571	.2589	.2595	.2591	.2602	.2609	.3218
								NDCG					
			@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	Mean
Mean			.1902	.2014	.2081	.2128	.2188	.2247	.2312	.2377	.2444	.2507	.3216
1T	M	F	.1547	.1737	.1834	.1915	.1984	.2045	.2118	.2179	.2244	.2303	.3059
1T	M	i.F	.1846	.1976	.2038	.2111	.2184	.2243	.2301	.2370	.2425	.2490	.3224
1T	M	s.F	.1800	.1956	.2019	.2091	.2148	.2207	.2266	.2333	.2385	.2450	.3179
1T	M	Fsum	.1884	.1954	.2054	.2128	.2218	.2281	.2336	.2396	.2457	.2515	.3222
MAXT	M	s.FumD	.1844	.2009	.2052	.2120	.2197	.2261	.2315	.2368	.2435	.2489	.3212
1T	Base	F	.1850	.1941	.1993	.2057	.2123	.2171	.2220	.2273	.2326	.2385	.3100
1T	Base	i.F	.1865	.1990	.2042	.2113	.2186	.2246	.2308	.2375	.2429	.2494	.3224
1T	Base	s.F	.1850	.1941	.1993	.2057	.2123	.2171	.2220	.2273	.2326	.2385	.3100
1T	Opt1	F	.1746	.1825	.1896	.1953	.1987	.2044	.2101	.2167	.2218	.2274	.2993
1T	Opt1	i.F	.1803	.1923	.1964	.2004	.2054	.2111	.2167	.2223	.2276	.2330	.3042
1T	Opt1	s.F	.1834	.1926	.1982	.2024	.2067	.2127	.2186	.2229	.2270	.2312	.3037
1T	Opt2	i.F	.1872	.1971	.2035	.2090	.2170	.2226	.2289	.2345	.2399	.2463	.3183
1T	Opt2	s.F	.1915	.2023	.2080	.2126	.2182	.2223	.2287	.2328	.2379	.2439	.3178
1T	Opt2	Fsum	.1906	.1978	.2039	.2119	.2193	.2245	.2308	.2359	.2417	.2473	.3184
MAXT	Opt2	Fsum	.1953	.1981	.2062	.2129	.2177	.2238	.2294	.2339	.2397	.2455	.3170

Table 4.9: 2008-agg: Precision and NDCG results of the BRE 1T for all the bba and distances evaluated with respect to the mean

Meth	BBA	W.	P@1	P@2	P@3	P@4	P@5	P@6	P@7	P@8	P@9	P@10	MAP
Mean			.2972	.3042	.2938	.2975	.2903	.2783	.2642	.2503	.2367	.2230	.3945
1T	M	F	.2436	.2659	.2725	.2775	.2745	.2676	.2547	.2419	.2277	.2164	.3738
1T	M	i.F	.1926	.2028	.2062	.2204	.2299	.2353	.2360	.2279	.2180	.2075	.3284
1T	M	s.F	.1391	.1971	.2270	.2443	.2518	.2544	.2524	.2407	.2283	.2168	.3239
1T	M	Fsum	.2551	.2869	.2883	.2883	.2816	.2719	.2593	.2433	.2297	.2173	.3832
MAXT	M	i.F	.1882	.2001	.2068	.2213	.2300	.2395	.2369	.2303	.2190	.2105	.3338
MAXT	M	Fsum	.2895	.2940	.2963	.2921	.2826	.2730	.2602	.2450	.2320	.2204	.3880
1T	Base	F	.2436	.2576	.2649	.2659	.2633	.2585	.2483	.2340	.2207	.2116	.362
1T	Base	i.F	.2092	.2130	.2219	.2258	.2298	.2322	.2312	.2231	.2127	.2043	.3334
1T	Base	s.F	.1455	.1939	.2156	.2318	.2375	.2449	.2421	.2334	.2208	.2097	.3132
1T	Opt1	F	.2347	.2455	.2572	.2602	.2561	.2540	.2451	.2321	.2208	.2101	.3536
1T	Opt1	i.F	.2679	.2749	.2738	.2746	.2666	.2626	.2496	.2368	.2245	.2133	.3734
1T	Opt1	s.F	.2475	.2570	.2670	.2666	.2638	.2589	.2483	.2352	.2236	.2134	.3548
1T	Opt2	F	.2564	.2768	.2760	.2819	.2763	.2691	.2560	.2417	.2270	.2155	.3775
1T	Opt2	i.F	.3074	.3029	.2976	.2937	.2834	.2749	.2622	.2493	.2341	.2213	.3973
1T	Opt2	s.F	.2781	.2768	.2866	.2873	.2808	.2723	.2635	.2503	.2344	.2222	.3771
1T	Opt2	Fsum	.2959	.2991	.2980	.2921	.2829	.2770	.2620	.2476	.2327	.2200	.3917
MAXT	Opt2	i.F	.3061	.3017	.2972	.2915	.2826	.2734	.2615	.2485	.2333	.2205	.3960
								NDCG					
			@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	Mean
Mean			.2368	.2806	.3080	.3432	.3713	.3888	.3992	.3724	.1643	.1694	.3895
1T	M	F	.1934	.2511	.2873	.3215	.3508	.3728	.3839	.3590	.1508	.1575	.3695
1T	M	i.F	.1654	.1971	.2208	.2545	.2848	.3133	.3337	.3111	.1265	.1336	.3172
1T	M	s.F	.1127	.1807	.2223	.2616	.2929	.3228	.3435	.3165	.1430	.1501	.3130
1T	M	Fsum	.2058	.2669	.3028	.3335	.3603	.3804	.3932	.3657	.1561	.1625	.3803
MAXT	M	i.F	.1616	.1951	.2231	.2575	.2874	.3204	.3373	.3134	.1236	.1323	.3176
MAXT	M	Fsum	.2360	.2744	.3074	.3362	.3622	.3832	.3941	.3671	.1583	.1665	.3830
1T	Base	F	.1926	.2383	.2736	.3041	.3327	.3561	.3689	.3421	.1399	.1477	.3534
1T	Base	s.F	.1174	.1751	.2094	.2453	.2739	.3065	.3262	.3030	.1323	.1393	.2977
1T	Base	i.F	.1760	.2097	.2380	.2637	.2896	.3144	.3353	.3134	.1214	.1292	.3217
1T	Opt1	F	.1743	.2204	.2572	.2893	.3164	.3413	.3543	.3290	.1341	.1415	.3394
1T	Opt1	i.F	.2117	.2544	.2849	.3152	.3397	.3642	.3755	.3489	.1441	.1507	.3638
1T	Opt1	s.F	.1871	.2286	.2664	.2966	.3263	.3477	.3598	.3336	.1390	.1477	.3460
1T	Opt2	F	.1986	.2576	.2886	.3249	.3535	.3742	.3855	.3596	.1518	.1582	.3710
1T	Opt2	s.F	.2169	.2520	.2893	.3231	.3498	.3722	.3867	.3577	.1622	.1680	.3696
1T	Opt2	i.F	.2461	.2846	.3124	.3422	.3670	.3880	.4005	.3742	.1620	.1672	.3904
1T	Opt2	Fsum	.2381	.2754	.3080	.3354	.3636	.3871	.3965	.3704	.1607	.1667	.3851
MAXT	Opt2	i.F	.2448	.2858	.3130	.3423	.3675	.3876	.4007	.3736	.1617	.1677	.3899

4.7.4 LETOR: Partioning the Data into Quality Clusters

We have partitioned the queries of each dataset in clusters that correspond to groups where the mean shows good, medium or bad performance. The goal is to analyze the performance of BRE and the mean in each group that corresponds to a different situation of quality of the true-rank estimator. This way to proceed is similar of what we have done on synthetic data where different quality cases (*good, equal and poor* cases) have been generated from the true ranking. Our expectation is that BRE should outperform the mean only in particular cases where the mean is a good estimator but not excellent, like in the total rankings experiments, where BRE shows some difficulties in the case *good* instead of the cases *poor* and *equal*. As measure on quality, we have decided to take the NDCG@Mean obtained by the mean. The choice of this measure is supported by the fact of taking in consideration the performance on the entire ranking and not only the top elements. Moreover, the $NDCG$ as evaluation measure is more important than the precision in the information retrieval literature. Starting from the median value of the NDCG@Mean obtained by the mean, we have split the queries in the following groups:

Bad Queries with NDCG@Mean values between the median and the $10-th$ percentile.

Good All the queries with NDCG@Mean values between the median and the $90-th$ percentile.

Tail Bad All the queries with NDCG@Mean values less than the $10-th$ percentile of the NDCG@Mean histogram.

Tail Good Queries with NDCG@Mean values greater than the $90-th$ percentile.

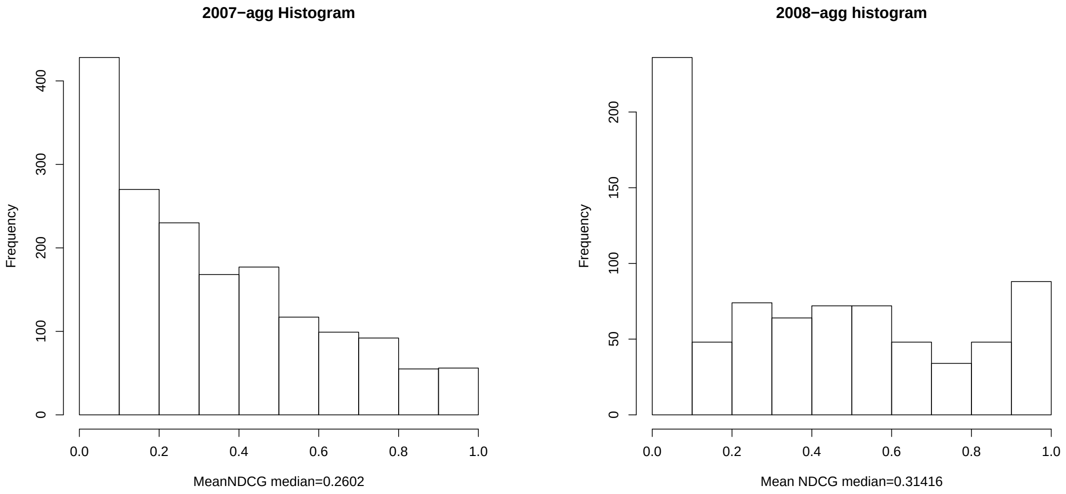
Moreover, we have joined all the queries that lies in the bad and good partitions in a unique partition called center. The center partition, using the median of the NDCG@Mean has been further split in two parts:

Center Bad All the queries with NDCG@Mean values less than the median of the NDCG@Mean in the center partition.

Center Good Queries with NDCG@Mean values greater than the median of the NDCG@Mean in the center partition.

From the histograms in both the datasets (Fig. 4.7) we notice a large number of queries where the mean shows value of NDGC@mean near to 0. The 2008-agg histogram (Fig. 4.7(b)) shows more queries in the two tails (left and right from the median) than the 2007 dataset where it presents a distribution more centered to tail-bad (left from the median). We have decided to not include all the combinations of bba and distance evaluated on BRE previously, but to focus only on the best settings like the M and Opt2 as BBA and $s.F$ and $i.F$ as distance. We have evaluated only the not-weighted and the 1T version due to the interesting results obtained in the previous experiments.

For both the datasets (Tab. 4.10-4.11), we notice that the mean shows NDCG values equals to 0 in all the queries in the tail-bad group. BRE cannot improve these queries due to the lack of any relation between the true-rank estimator and the true ranking. This adds another difficulty to improving the mean, since this group contains a large part



(a) Histogram. 2007-aggr

(b) Histogram. 2008-aggr

Figure 4.7: Histogram of NDCG@Mean obtained by the median in the 2007-agg and 2008-agg datasets.

of the queries in both datasets. In Fig. 4.8 and Fig. 4.9 are plotted the best results of BRE-1T with respect to the mean for all the clusters of queries evaluated.

From Fig. 4.8 we notice that BRE outperforms the mean for the queries that belongs in the bad and in the center-bad groups whereas mean obtains better performance in the tail-good, good and the center groups. We highlight that the best results of BRE in the good and bad group, are founded using different bba. In the bad group, the bba M shows the best result instead of the Opt2 bba in the tail-group. In the center group that includes the bad and good ones, BRE outperforms the mean with bba M and $i.F$ distance. As our expectation the same configuration of BRE improves with large margin the mean in the center-bad partition, and obtains NDCG values really similar to the mean in center-good. Moreover, BRE with the bba M shows respectively .0500 and .0701 at NDCG@1 in the bad and center-bad partition where the mean obtains really poor performances (.0028, .0042 NDCG@1).

Regarding to the 2008-agg dataset, from Fig. 4.9 we notice that BRE with respect to the mean follows what we have seen in the previous dataset. BRE outperforms the mean with low margin in the bad group and shows competitive results in the good partitions (Fig. 4.9). As consequence, BRE outperforms the mean also in the center partition in almost all the NDCG values. Dividing the center group into center-bad and center good, we notice that BRE has the most effective performance in the center-bad instead of the center-good in which the mean has slight improvement. As showed in Fig. 4.7(b), the high NDCG values of BRE in the good partitions are also caused by the low number of queries in that partition than the correspondent partition in the 2007-agg dataset. In both the tail-good partitions the BRE does not improve the mean. With respect to 2007-agg dataset, the bba Opt2 gives the best belief assignment due to the presence of high partiality in the queries of the 2008-agg dataset.

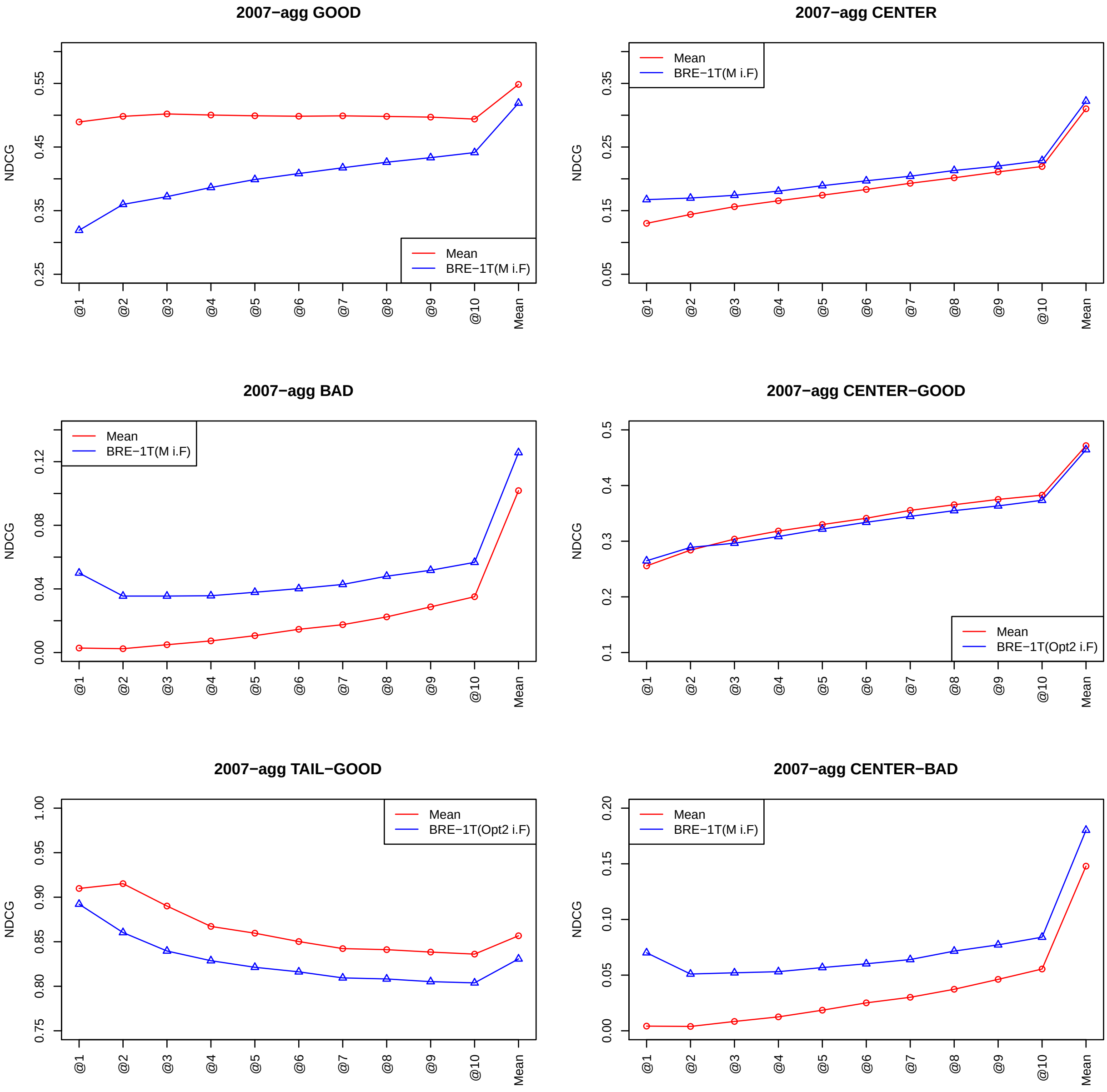


Figure 4.8: 2007-agg: Best results of BRE-1T with respect to the mean for all the clusters evaluated

In this experiment, we give an empirical explanation of why the BRE algorithm does not provide a solid improvement with respect to the mean in the two LETOR datasets. However the difficulties of the LETOR task in terms of partial true-rankings, BRE with a wide range of different bba evaluated, improves the mean only in that queries where the mean performance are not very good. For a large part of the queries the mean is not a useful estimator method, as a consequence BRE can not improve this situation.

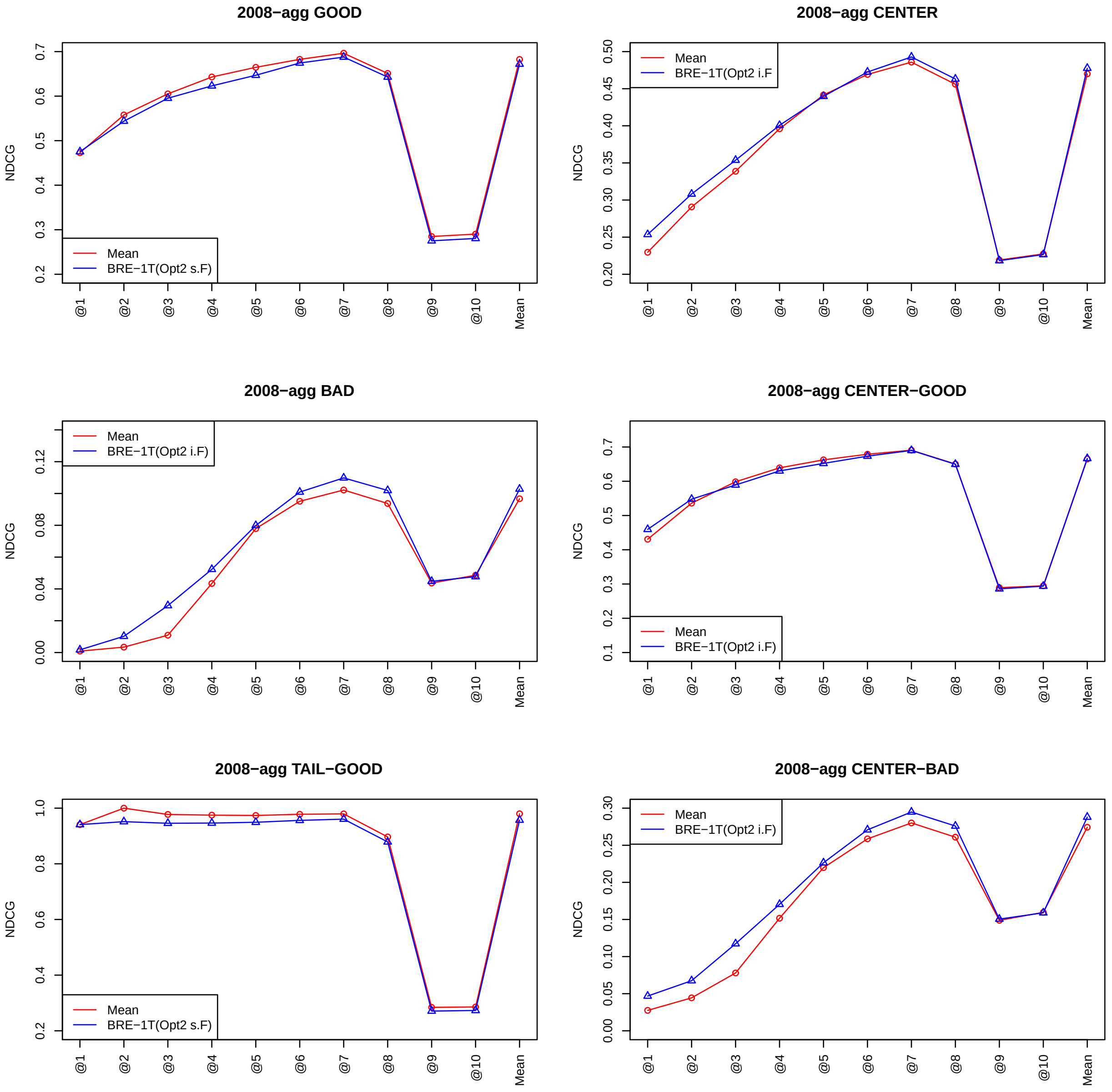


Figure 4.9: 2008-agg: Best results of BRE-1T with respect to the mean for all the clusters evaluated

Moreover this overall overview of the BRE performance in this task corresponds quite well to the advantages of BRE on cases of heterogeneous qualities of the input rankings. From the not so brilliant results of BRE in the good partitions, several causes of why this happen can be discussed but it will be deeply analyzed in future works. As in the previous experiments, the important role of the bba and how they could fit the different nature of the rankings to aggregate are further highlighted.

Table 4.10: LETOR 2007-agg: Comparison of BRE and the mean in terms of NDCG on all the partions evaluated

							Good						
Meth.	BBA	W.	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	Mean
Mean			.4894	.4982	.5020	.5003	.4991	.4984	.4990	.4981	.4970	.4939	.5484
NW	Opt2		.3243	.3492	.3564	.3656	.3754	.3840	.3927	.4004	.4078	.4167	.5004
1T	Opt2	i.F	.3668	.3845	.3957	.4037	.4154	.4225	.4315	.4390	.4460	.4527	.5300
1T	Opt2	s.F	.3739	.3847	.3935	.4049	.4145	.4217	.4304	.4368	.4445	.4513	.5283
1T	M	i.F	.3191	.3598	.3720	.3864	.3989	.4083	.4174	.4261	.4333	.4413	.5191
							Bad						
Mean			.0028	.0024	.0049	.0073	.0106	.0146	.0175	.0224	.0287	.0351	.1018
NW	Opt2		.0142	.0112	.0135	.0161	.0184	.0226	.0247	.0294	.0336	.0396	.1039
1T	Opt2	i.F	.0075	.0097	.0113	.0143	.0186	.0225	.0262	.0300	.0337	.0398	.1066
1T	Opt2	s.F	.0071	.0108	.0141	.0189	.0240	.0272	.0311	.0349	.0390	.0433	.1084
1T	M	i.F	.0500	.0355	.0355	.0357	.0379	.0402	.0428	.0480	.0517	.0567	.1257
							Tail-Good						
Mean			.9098	.9152	.8901	.8672	.8596	.8502	.8423	.8411	.8384	.8361	.8567
NW	opt		.7412	.7598	.7580	.7525	.7487	.7490	.7499	.7488	.7473	.7473	.7836
1T	Opt2	i.F	.8392	.8387	.8285	.8208	.8160	.8124	.8076	.8036	.8010	.8007	.8262
1T	Opt2	s.F	.8922	.8603	.8396	.8287	.8213	.8162	.8094	.8082	.8052	.8038	.8307
1T	M	i.F	.5725	.6833	.7118	.7353	.7438	.7434	.7467	.7480	.7497	.7499	.7731
							Tail-Bad						
Mean	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000		
							Center						
Mean			.1300	.1440	.1562	.1655	.1743	.1833	.1930	.2016	.2109	.2194	.3101
NW	Opt2		.1248	.1368	.1432	.1517	.1602	.1686	.1756	.1839	.1917	.2015	.2942
1T	Opt2	i.F	.1354	.1485	.1583	.1666	.1778	.1855	.1945	.2024	.2099	.2184	.3098
1T	Opt2	s.F	.1328	.1466	.1573	.1694	.1801	.1875	.1968	.2037	.2118	.2193	.3093
1T	M	i.F	.1673	.1698	.1741	.1806	.1892	.1969	.2041	.2131	.2201	.2287	.3222
							Center-Good						
Mean			.2556	.2840	.3038	.3183	.3299	.3412	.3554	.3655	.3751	.3827	.4717
NW	Opt2		.2253	.2551	.2643	.2773	.2905	.3009	.3120	.3217	.3314	.3426	.4376
1T	Opt2	i.F	.2586	.2814	.2983	.3099	.3253	.3351	.3476	.3576	.3671	.3753	.4652
1T	Opt2	s.F	.2648	.2889	.2964	.3084	.3218	.3340	.3446	.3549	.3634	.3735	.4644
1T	M	i.F	.2648	.2889	.2964	.3084	.3218	.3340	.3446	.3549	.3634	.3735	.4644
							Center-Bad						
Mean			.0042	.0039	.0084	.0125	.0185	.0251	.0301	.0373	.0462	.0555	.1479
NW	Opt2		.0244	.0187	.0223	.0264	.0301	.0365	.0394	.0464	.0524	.0608	.1510
1T	Opt2	i.F	.0125	.0158	.0186	.0235	.0305	.0363	.0418	.0475	.0529	.0617	.1545
1T	Opt2	s.F	.0114	.0171	.0225	.0306	.0383	.0435	.0487	.0540	.0601	.0661	.1571
1T	M	i.F	.0701	.0510	.0521	.0531	.0568	.0602	.0640	.0716	.0772	.0841	.1802

Table 4.11: LETOR 2008-agg: Comparison of BRE and the mean in terms of NDCG on all the partions evaluated.

	@1	@2	@3	@4	@5	@6	@7	@8	@9	@10	Mean
Good											
Mean	.4728	.5578	.6051	.6430	.6647	.6826	.6963	.6511	.2848	.2902	.6824
NW Opt2	.3486	.4826	.5379	.5780	.6034	.6282	.6460	.6078	.2588	.2662	.6182
1T Opt2 i.F	.4906	.5591	.5952	.6321	.6541	.6751	.6912	.6465	.2792	.2868	.6781
1T Opt2 s.F	.4753	.5438	.5953	.6230	.6469	.6744	.6877	.6428	.2751	.2804	.6719
1T M i.F	.2934	.3550	.3934	.4497	.4944	.5303	.5582	.5217	.2131	.2222	.5257
Bad											
Mean	.0009	.0034	.0109	.0434	.0779	.0951	.1022	.0937	.0437	.0486	.0967
NW Opt2	.0094	.0191	.0274	.0574	.0818	.1010	.1100	.0982	.0441	.0503	.1042
1T Opt2 i.F	.0017	.0102	.0295	.0523	.0799	.1009	.1098	.1019	.0448	.0477	.1028
1T Opt2 s.F	.0009	.0070	.0208	.0478	.0804	.0998	.1053	.0980	.0463	.0528	.0983
1T M i.F	.0374	.0393	.0481	.0593	.0752	.0964	.1092	.1005	.0399	.0451	.1088
Tail-Good											
Mean	.9409	1.0000	.9774	.9749	.9737	.9780	.9795	.8966	.2843	.2856	.9799
NW opt	.7975	.7342	.6540	.5823	.5165	.4705	.4159	.3734	.3376	.3089	.8747
1T Opt2 i.F	.8861	.9335	.9285	.9358	.9429	.9498	.9493	.8708	.2672	.2692	.9426
1T Opt2 s.F	.9409	.9515	.9457	.9464	.9493	.9561	.9605	.8787	.2709	.2730	.9569
1T M i.F	.3924	.5105	.5714	.6416	.6868	.7226	.7429	.6804	.1894	.1903	.6531
Tail-Bad											
Mean	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
Center											
Mean	.2296	.2907	.3387	.3960	.4416	.4692	.4858	.4560	.2192	.2274	.4700
NW Opt2	.2433	.3124	.3320	.3531	.3493	.3471	.3390	.3265	.3111	.2975	.4510
1T Opt2 i.F	.2536	.3081	.3537	.4007	.4397	.4725	.4928	.4631	.2183	.2265	.4776
1T Opt2 s.F	.2316	.2902	.3439	.3880	.4332	.4700	.4845	.4556	.2157	.2249	.4667
1T M i.F	.2034	.2356	.2638	.3069	.3486	.3888	.4185	.3921	.1737	.1851	.4065
Center-Good											
Mean	.4307	.5360	.5985	.6392	.6625	.6789	.6907	.6502	.2894	.2948	.6649
NW Opt2	.2826	.4448	.5100	.5622	.5937	.6175	.6340	.6004	.2638	.2700	.5903
1T Opt2 i.F	.4595	.5477	.5891	.6301	.6521	.6733	.6900	.6496	.2859	.2937	.6665
1T Opt2 s.F	.4348	.5329	.6001	.6338	.6522	.6792	.6893	.6494	.2872	.2914	.6626
1T M i.F	.2963	.3426	.3839	.4433	.4883	.5272	.5541	.5221	.2167	.2256	.5190
Center-Bad											
Mean	.0275	.0444	.0779	.1517	.2198	.2586	.2801	.2609	.1488	.1596	.2743
NW Opt2	.0413	.0706	.1032	.1621	.2109	.2542	.2806	.2584	.1364	.1516	.2787
1T Opt2 i.F	.0468	.0675	.1173	.1704	.2264	.2708	.2948	.2759	.1505	.1590	.2879
1T Opt2 s.F	.0275	.0465	.0865	.1413	.2134	.2599	.2788	.2611	.1439	.1581	.2700
1T M i.F	.1102	.1281	.1432	.1700	.2083	.2498	.2822	.2616	.1304	.1444	.2935

4.8 Conclusions

In this chapter we have applied the Belief Ranking Estimator (BRE) on the aggregation of partial rankings. The partial rankings and the top-k rankings are met in most of the real problems where combined rankings are involved . The introduction of the partial rankings makes the task of the estimation of the true ranking quite difficult, due to lack of any relation to the true ranking underlying the problem. Through the use of synthetic data we have evaluated a straightforward version of BRE against the competitors on three cases of partial rankings that correspond to different hypothesis on their generation.

In the first experiment, BRE shows interesting results with respect to the competitors when the top-k rankings are generated from experts that share all the items in the universe set. We point out that the iterative schema gives a considerable improvement also with short rankings. The performance of BRE and the competitors decrease when the partial rankings are generated by experts that do not have a complete knowledge of the universe set, where as consequence also the relation to the true ranking is weak. Despite of the difficulty in these two experiments, BRE outperformed significantly some competitors also when the rankings have in common very few items. From the experiments on the synthetics data BRE gets interesting preliminary results that show the great flexibility of our method to estimate a true ranking dealing with partial rankings. On the other hand, the major drawbacks highlighted are the (1)quality of true-rank estimator used and its relation to the true ranking and (2) the possibility to deal with rankings with very few items in common.

From the synthetic data results, we have decided to apply BRE on the LETOR benchmark where the task is to combine rankings from search engines retrieved on different queries. On LETOR datasets, the comparison of BRE with respect to the mean of the rankings showed interesting results on this real data task but not enough to strongly outperform the mean. On the other hand, on LETOR we have shown the role of the belief assignment inside BRE, since belief assignment that measures the partiality of the input rankings works well instead of the belief assignment used in the total rankings. Also for the distance used for the weights computation, distance designed on partial rankings such as the induced and the scaled Spearman footrule distance have improved the results instead of the standard Spearman footrule distance. A further deep investigation on this dataset has showed that BRE outperforms with a large margin the mean, only in queries where the mean does not have so very good performance. Unfortunately, BRE does not improve the queries where the mean shows very good performance.

Even if the performance of BRE on partial rankings are not supported by impressive results with respect to the competitors, we have showed the flexibility of BRE, given by the Belief function, to model *a priori* knowledge of the problem that lead to improve the estimation of the true ranking. The work on the aggregations of partial rankings is not concluded with this experiments, the role of the true-rank estimator together with a local weight schema should be investigated in future works.

Chapter 5

Conclusions

In this work we have faced the ranking aggregation problem, taking into account the true ranking in its formulation. As solution we propose an algorithm called Belief Ranking Estimator (BRE), that in an unsupervised way estimates the true ranking given a set input rankings. Through the use of the belief function framework, we model the uncertainty of each rank and combine them accordingly to weights computed as distances from a true-rank estimator that can be provided by a ranking aggregation method. We have proposed and evaluated three versions of BRE. The not weighted version in which the rankings are aggregated using only the belief derived from the input rankings without the application of the weights. The second version is the iterative one where the belief distribution of the rankings are discounted by weights, and at each step the less informative input ranking is replaced by the combined ranking. In case of only one step, we refer to the third version of BRE as the weighted version. We have evaluated BRE on the aggregation of total and partial rankings, comparing the results against some state-of-the-art methods.

On total rankings we have generated an experimental setting based on synthetic data where the input rankings show diverse quality with respect to the true ranking. We have compared the BRE performance with respect to the mean, median and the optimal footrule aggregation, that have been also used as true-rank estimator. BRE with the weighting schema has showed quite impressive results with respect to the competitors evaluated in all the cases with low-quality rankings. We point out that BRE with weighting schema outperforms the true-rank estimator for many competitors such as the median and the mean. The performance of BRE seems to be not affected by the number of rankings, since also with 30 input rankings the weighting schema shows its superiority to fit the real quality of the input rankings. To explore the limits of our solution we have aggregated rankings that are really similar and consequently highly informative, to the true ranking. As we expected, we found that BRE suffers in comparison to the mean, in case of extremely homogeneous-quality rankings. Moreover, we have explored through several experiments the role of the weights, and of the weighting schemas.
With regard to the quality of the weights, we have proposed and evaluated Quality BRE (QBRE), a novel algorithm that aims to find the best weights of the input rankings. The comparison of the weights obtained with QBRE against BRE's weights, has showed the high quality of the weights computed by QBRE. Finally, the weights provided by QBRE

has been evaluated on BRE, showing that the use of better weights improves the performance significantly. We have also proposed different weighting schemas on BRE, but the results showed that the simple schema is better.

Moving to the aggregation of the partial and top-k rankings, BRE has met more difficulties to outperform the competitor methods. We point out that on this type of rankings we have used the straightforward BRE used on total rankings, except for some minor modifications strictly needed to deal with rankings that do not have the same items. Using synthetic data we have generated three different cases of generation partial/top-k lists on which we have evaluated BRE and some competitor methods. In the first case, where the top-k rankings are generated from experts that share all the items in the universe set, BRE still defends its good results with respect to the competitor methods. In the other two cases, where the partial rankings are generated by experts that do not have a total knowledge of the universe set, BRE and the competitor methods show bad performance. From the synthetic data results, we have realized how the aggregation of partial rankings is a more complex situation for the following reasons: (1) the possibility to aggregate rankings with few items in common (2) the weak correlation of the true-rank estimator with the true rankings.
With respect to the evaluation on real data, we have applied BRE on the LETOR datasets where we have used the mean as competitor and true-rank estimator. BRE showed interesting results but not good enough to totally outperform the mean. Despite the not so brilliant results of BRE on this specific task, we have evaluated several bbas showing how the bbas that take into account the partiality of the input rankings work better with respect to the bbas used for the total rankings. Moreover, we have showed that distances designed for partial rankings, such as the induced and scaled Spearman footrule distance improve the results.

We point out that this is the first work that uses the belief function to combine rankings, so several questions are still open and need to be deeply investigated. Although the experimental results provided in this work cover just a part of the possible questions, we can conclude that our algorithm is an effective solution to estimate the true ranking. From the results on total rankings, BRE outperforms the competitors in particular in the situation where the input rankings show heterogeneous quality with respect to the true ranking. The weighting schema and the weight computation are the key points of our approach, and we have showed how good weights increase the performance of BRE. As future work we want to evaluate BRE on real tasks where total rankings are involved in order to assess the performance of BRE and QBRE using other distances such as the Kendall distance. Although the result on the partial rankings show some difficulties for BRE to outperform the competitors, we have showed that the possibility to encode different *a priori* information by many different bbas. The major difficulty that we need to overcome is the combination of disjoint set of items and the totally-unknown relation of the input rankings with respect to true ranking. As future work we want to improve the combination of partial/top-k rankings, investigating local weighting schemas where the local weights are applied to single items instead of global weights applied to all the ranked elements.

A preliminary version of the results showed in Chapter 3 has been presented as student poster at ECSQARU 2011 [1]. A more complete version of Chapter 3 is currently submitted at IPMU 2011 [2]. During my Ph.D. I also focused my attention on learning algorithms such as kernel methods and Nearest Neighbours (k-NN) classifiers. With regard to the distance used in the K-NN classifier, I have performed an empirical comparison of two distance metrics presented in the literature [3].

Bibliography

[1] S. Lin, "Rank Aggregation Methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 555–570, 2010.

[2] S. Lin and J. Ding, "Integration of Ranked Lists via Cross Entropy Monte Carlo with Applications to mRNA and microRNA Studies," *Biometrics*, vol. 65, no. 1, pp. 9–18, 2009.

[3] T. Qin, T. Liu, J. Xu, and H. Li, "LETOR: A Benchmark Collection For Research on Learning to Rank for Information Retrieval," *Information Retrieval*, pp. 1–29, 2010.

[4] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank Aggregation Methods for the Web," in *Proceedings of the 10-th WWW Conference*, pp. 613–622, 2001.

[5] A. Klementiev, D. Roth, and K. Small, "Unsupervised Rank Aggregation with Distance-Based Models," in *ICML'08: Proceedings of the 25th International Conference on Machine learning*, pp. 472–479, 2008.

[6] P. Diaconis and R. L. Graham, "Spearman's Footrule as a Measure of Disarray," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 30, no. 2, pp. 262–268, 1979.

[7] L. Thurstone and L. Jones, "The Rational Origin for Measuring Subjective Values," *Journal of the American Statistical Association*, pp. 458–471, 1957.

[8] L. Thurstone, "Rank Order as a Psycho-Physical Method," *Journal of Experimental Psychology*, vol. 14, no. 3, p. 187, 1931.

[9] R. Luce, *Individual Choice Behavior.* John Wiley, 1959.

[10] R. Plackett, "The Analysis of Permutatioos," *Applied Statistics*, vol. 24, no. 2, pp. 193–202, 1975.

[11] C. Mallows, "Non-null Ranking Models," *Biometrika*, vol. 44, no. 1/2, pp. 114–130, 1957.

[12] D. Critchlow, *Metric Methods for Analyzing Partially Ranked Data*, vol. 34. Springer, 1985.

[13] G. Lebanon and J. Lafferty, "Conditional Models on the Ranking Poset," *Advances in Neural Information Processing Systems*, vol. 15, pp. 415–422, 2002.

[14] V. Pihur, S. Datta, and S. Datta, “Weighted Rank Rggregation of Cluster Validation Measures: a Monte Carlo Cross-Entropy Approach,” *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, 2007.

[15] V. Pihur, S. Datta, and S. Datta, “RankAggreg, An R Package for Weighted Rank Aggregation,” *BMC Bioinformatics*, vol. 10, no. 1, p. 62, 2009.

[16] J. Borda, *Mémoire sur les Élections au Scrutin.* Histoire de l’ Académie Royale des Sciences, 1781.

[17] R. Fagin, R. Kumar, and D. Sivakumar, “Efficient Similarity Search and Classification via Rank Aggregation,” in *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data*, pp. 301–312, 2003.

[18] A. Klementiev, D. Roth, and K. Small, “An unsupervised learning algorithm for rank aggregation,” in *Machine Learning: ECML 2007*, vol. 4701, pp. 616–623, 2007.

[19] T. Qin, X. Geng, and T.-Y. Liu, “A New Probabilistic Model for Rank Aggregation,” in *Advances in Neural Information Processing Systems 23, NIPS 2010*, pp. 1948–1956, 2010.

[20] M. Kendall, *Rank Correlation Methods.* Griffin, 1948.

[21] R. Fagin, R. Kumar, and D. Sivakumar, “Comparing top k Lists,” in *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 28–36, 2003.

[22] J. Myers and A. Well, *Research Design and Statistical Analysis*, vol. 1. Lawrence Erlbaum, 2003.

[23] H. Young, “Condorcet’s Theory of Voting,” *The American Political Science Review*, pp. 1231–1244, 1988.

[24] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, “An Efficient Boosting Algorithm for Combining Preferences,” *J. Mach. Learn. Res.*, vol. 4, pp. 933–969, 2003.

[25] R. Herbrich, T. Graepel, and K. Obermayer, “Large Margin Rank Boundaries for Ordinal Regression,” *Advances in Large Margin Classifiers*, vol. 88, no. 2, pp. 115–132, 2000.

[26] T. Liu, “Learning to Rank for Information Retrieval,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 3, pp. 225–331, 2009.

[27] S. Adali, B. Hill, and M. Magdon-Ismail, “The Impact of Ranker Quality on Rank Aggregation Algorithms: Information vs. Robustness,” in *Proceedings of the 22nd International Conference on Data Engineering*, pp. 37–37, 2006.

[28] P. Alexiou, M. Maragkakis, G. Papadopoulos, M. Reczko, and A. Hatzigeorgiou, “Lost in Translation: An Assessment and Perspective for Computational microRNA Target Identification,” *Bioinformatics*, vol. 25, no. 23, pp. 3049–55, 2009.

[29] A. P. Dempster, “Upper and Lower Probabilities Generated by a Random Closed Intervals,” *The Annals of Mathematical Statistics*, vol. 39, pp. 957–966, 1968.

[30] G. Shafer, *A Mathematical Theory of Evidence*, vol. 1. Princeton University press Princeton, 1976.

[31] P. Smets and R. Kennes, “The Transferable Belief Model,” *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, 1994.

[32] J. Schubert, “Clustering Decomposed Belief Functions Using Generalized Weights of Conflict,” *International Journal of Approximate Reasoning*, vol. 48, no. 2, pp. 466–480, 2008.

[33] D. Dubois and H. Prade, “Representation and Combination of Uncertainty with Belief Functions and Possibility Measures,” *Computational Intelligence*, vol. 4, no. 3, pp. 244–264, 1988.

[34] R. Yager, “On the Dempster-Shafer Framework and New Combination Rules,” *Information sciences*, vol. 41, no. 2, pp. 93–137, 1987.

[35] T. Denceux, “The Cautious Rule of Combination for Belief Functions and Some Extensions,” in *Information Fusion, 2006 9th International Conference on*, pp. 1–8, IEEE, 2006.

[36] L. Oukhellou, A. Debiolles, T. Denœux, and P. Aknin, “Fault Diagnosis in Railway Track Circuits Using Dempster-Shafer Classifier Fusion,” *Engineering Applications of Artificial Intelligence*, vol. 23, no. 1, pp. 117–128, 2010.

[37] P. Smets, “Decision Making in the TBM: the Necessity of the Pignistic Transformation,” *International Journal of Approximate Reasoning*, vol. 38, no. 2, pp. 133–147, 2005.

[38] P. Smets, “Decision Making in the TBM: The Necessity of the Pignistic Transformation,” *International Journal of Approximate Reasoning*, vol. 38, no. 2, pp. 133–147, 2005.

[39] B. Cobb and P. Shenoy, “On the pPlausibility Transformation Method for Translating Belief Function mModels to Probability Models,” *International Journal of Approximate Reasoning*, vol. 41, no. 3, pp. 314–330, 2006.

[40] T. Denœux and P. Smets, “Classification Using Belief Functions: Relationship Between Case-based and Model-Based Approaches,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 36, no. 6, pp. 1395–1406, 2006.

[41] R. Srivastava and G. Shafer, “Belief Function Formulas for Audit Risk,” *Classic Works of the Dempster-Shafer Theory of Belief Functions*, pp. 577–618, 2008.

[42] R. Yager, “Decision Making Under Dempster-Shafer Uncertainties,” *International Journal Of General System*, vol. 20, no. 3, pp. 233–245, 1992.

[43] L. Zouhal and T. Denœux, "An Evidence-Theoretic k-NN Rule with Parameter Optimization," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 28, no. 2, pp. 263–271, 1998.

[44] Z. Younes, F. Abdallah, and T. Denœux, "An Evidence-Theoretic K-Nearest Neighbor Rule for Multi-Label Classification," *Scalable Uncertainty Management*, pp. 297–308, 2009.

[45] M. Masson and T. Denœux, "ECM: An Evidential Version of the Fuzzy C-Means Algorithm," *Pattern Recognition*, vol. 41, no. 4, pp. 1384–1397, 2008.

[46] T. Denœux and M. Masson, "EVCLUS: Evidential Clustering of Proximity Data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 34, no. 1, pp. 95–109, 2004.

[47] L. Serir, E. Ramasso, and N. Zerhouni, "E2GK:Evidential Evolving Gustafsson-Kessel Algorithm for Data Streams Partitioning Using Belief Functions," in *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pp. 326–337, 2011.

[48] T. Denœux and M. Masson, "Dempster-Shafer Reasoning in Large Partially Ordered Sets: Applications in Machine Learning," *Integrated Uncertainty Management and Applications*, pp. 39–54, 2010.

[49] R. Lo Cigno, A. Russo, and D. Carra, "On Some Fundamental Properties of P2P Push/Pull Protocols," in *Communications and Electronics, 2008. ICCE 2008. Second International Conference on*, pp. 67–73, 2008.

[50] K. Järvelin and J. Kekäläinen, "IR Evaluation Methods for Retrieving Highly Relevant Documents," in *Proceedings of SIGIR 2000*, pp. 41–48, 2000.

[51] K. Järvelin and J. Kekäläinen, "Cumulated Gain-Based Evaluation of IR Techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

Appendix A

List of Publications

[1] A. Argentini and E. Blanzieri, “Unsupervised Learning of True Ranking Estimators using the Belief Function Framework,” Tech. Rep. DISI-11-480, University of Trento, November 2011.

[2] A. Argentini and E. Blanzieri, “Ranking Aggregation Based on Belief Function.” Submitted to IPMU, 2012.

[3] A. Argentini and E. Blanzieri, “About Neighborhood Counting Measure Metric and Minimum Risk Metric,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 763–765, 2010.