ATLANTIS PRESS

Advances in Computer Science Research, volume 87

3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)

Comparing Rank Aggregation Methods based on Mallows Model

Zhangqian Zhu [a], Xiaomeng Wang [b, *] and Shigang Qiu [c]

College of Computer & Information Science, Southwest University, Beibei 400715, China.

[a]zhangqianzhu1994@163.com, [b, *] wxm1706@swu.edu.cn, [c]qsg6038@163.com

Abstract. Rank aggregation is the process of aggregating multiple base rankers into a single but more comprehensive ranker, which plays an important role in many domains such as recommender system, meta-search, database, genomics, etc. Works related to the comparison of rank aggregation methods all don't have a suitable and general data generation mechanism to produce data with various characteristics and lack a more reasonable and effective algorithm evaluation performance index. Therefore, this paper presents a general data generation mechanism based on Mallows model to produce synthetic controllable datasets, uses generalized Kendall rank correlation coefficient and rank-biased overlap to evaluate and compare the performance of two kinds of methods under different settings. Besides, we also consider the comparison between indices and the impact of data characteristics on the algorithms. This paper may be helpful to researchers and decision-makers from multiple domains.

Keywords: Rank aggregation; Mallows model; rank-biased overlap.

1. Introduction

Rank aggregation, also known as Kemeny rank aggregation [1], preference aggregation [2], consensus ranking problem [3], aims at aggregating multiple base rankers into a single but more comprehensive ranker, which is considered to be more reliable and trustworthy than baser rankers, and plays an important role in many domains such as recommender system [4], meta-search [5], database, genomics, etc. It has a rich research history with more than two hundred years which dates back to work of Jean-Charles de Borda in 1770 [6]. Based on the basic idea that "the whole is greater than the sum of its parts", a large number of rank aggregation methods have been proposed from different research fields, such as social choice and mathematics, statistics, genomics, information retrieve, which can be roughly divided into ad hoc methods and distance-based methods [7]. The former combines multiple ranking lists into one based on some kind of intuition or man-made rule, which is simple and fast, while the latter seeks a consensus that is defined to be that set of preferences which is closest, in a minimum distance or maximum correlation coefficient sense, to base ranker responses, and appears time-consuming. Because finding such a consensus is NP-hard, it's difficult to guarantee an optimal solution even though distance-based methods is more complete in theory.

Comparing and analyzing different rank aggregation methods can help us to select the most suitable rank aggregation methods under different application scenarios and data conditions, and improve work efficiency, reliability, and credibility. For example, results from peer review of research proposals and articles, an essential element in R&D process and the academic community worldwide, can be combined to achieve better set of candidate proposals and help improve quality of decision output [8]. Although there are works [9-12] comparing rank aggregation methods from different aspects, all of them didn't have a suitable, consistent and general data generation mechanism. As far as we know, [9] is the first one to consider data generation model to produce sets of permutations with various degree of consensus, with focus on balance between search time and algorithm performance, but they didn't consider other list types. Furthermore, they all considered traditional indices to evaluate and compare the performance of rank aggregation methods and didn't realize that those indices are problematic more or less, which will be discussed in section 4. [10] developed a data generation model which can generate the required synthetic base rankers with adjustable accuracy and length and found both the accuracy and length have a remarkable effect on the comparison results between rank aggregation methods. However, their model is not controllable

to some extent, such as the number of ties, and their performance index is also not applicable since ground truth ranking is not easy to find and it is exactly what we are pursuing.

Therefore, this paper presents a general data generation mechanism based on Mallows model to produce synthetic controllable datasets, uses generalized Kendall rank correlation coefficient and rank-biased overlap to evaluate and compare the performance of two kinds of methods under different settings. Besides, we also consider the comparison between indices and the impact of data characteristics on the algorithms.

The rest of the paper is organized as follows: in the second section, we present an introduction of bucket order, list type and Mallows model on which our data generation mechanism is based. In the third section, we introduce two algorithms: PageRank and FAST, representing ad hoc method and distance-based method. In the fifth section, we describe two performance evaluation indices: generalized Kendall rank correlation coefficient and rank-biased overlap. In the sixth part, results and discussion will be conducted. The last part is the summary.

2. Data Generation Mechanism

Before the introduction of the Mallows model, we first introduce some basic concepts to avoid ambiguity.

Bucket order [13]: given set S containing |S| items, a bucket order is a transitive binary relation $\prec$ where there are $B_1, \ldots, B_t$ ($1 \le t \le |S|$) that form a partition of S such that item $x \prec y$ if and only if there are i, j with $i < j$ such $x \in B_i$ and $y \in B_j$. If $x \in B_i$, we regard B_i as the bucket of x and B_i precedes B_j if $i < j$. Intuitively, a bucket order appears a strict linear order with ties. After the definition of bucket order, all bucket orders can be divided into 3 types according to the status of an item in a list, number of items and lists:

Full linear ranking (FLR): size of all buckets in a bucket order is 1 and $t = |S|$ (e.g., all items contained and tie(s) not permitted, also called permutation).

Weak linear ranking (WLR): there is at least one bucket whose size is more than 1 and $1 \le t < |S|$ (e.g., the number of all items contained equals to |S| and tie(s) permitted).

Incomplete ranking (IR): bucket order whose number of items contained is less than |S| and there may have tie(s). IR_μ is regarded as an incomplete ranking under the ranking mechanism μ to distinguish it from others under different ranking mechanisms. When there exists N incomplete rankings $IR_{\mu_1}, \ldots, IR_{\mu_N}$ ($|IR_{\mu_i}| = l_i$, $i = 1, \ldots, N$), l_i is not necessarily identical.

In order to evaluate and compare rank aggregation methods, it's necessary to generate synthetic datasets with different settings. Since WLR and IR can be seen as variants of FLR, so we first generate data of FLR type. There are many generation models for FLR type data. To explore the impact of the degree of consensus from data on the algorithm and make data generation process controllable, the Mallows model [14] will be used in this paper. The Mallows model is an exponential location model, usually considered as analogous to the Gaussian distribution for permutations, and contains two parameters: (1) central permutation σ_0, give the mode of distribution and the probability of any other permutation increases as we move "closer" to the central permutation; (2) dispersion parameter, θ, controls how fast this increase happens. The Mallows model assigns each permutation with a probability value

$$p(\sigma|\theta, \sigma_0) = \frac{1}{Z(\theta)} e^{-\theta d(\sigma, \sigma_0)} \quad (1)$$

where $Z(\theta) = \sum_\sigma e^{-\theta d(\sigma, \sigma_0)}$ is normalization constant and d is some kind of distance, representing closeness between the central and any other permutations, such as Kendall τ distance, Hamming distance, Cayley distance, and Ulam distance. We pick Kendall τ distance in this paper.

After the generation of FLR type data, we can get other types of data by proper transformation from FLR data as follows:

(1) FLR → WLR

According to the difference between the definition of FLR and WLR, transformation to WLR requires to add ties to FLR. There are two rules to follow: first, number of ties add to FLR for all base lists $T \sim U(0,\ r_t * M)$, where M is the number of items and r_t is tie ratio between number of ties and M; Second, for all base lists, the number of items contained in a tie increases linearly from top to bottom of FLR (e.g., tie is more likely to happen on the end than on the top).

(2) $\text{WLR} \xrightarrow{k} \text{IR}$

Given base length k, length of all IR lists $L \sim U(k - \Delta k,\ k + \Delta k)$, where Δk determines both lower and upper bound of IR length. Especially, when $\Delta k = 0$, all IR lists have identical length, which is often called top k list in other literature.

All kinds of lists can be generated under the above data generation mechanism, and our experimental parameter settings is as follows: number of lists $N \in [10, 30]$ with a step of 5; number of items $M \in [20, 100]$ with a step of 20; degree of consensus $\theta \in [0.001, 0.01, 0.1, 0.4, 0.7]$ representing data quality from none of consensus to strong consensus; $r_k \in [0.5,\ 0.9]$ with a step of 0.1, representing ratio between base length k and M; $r_t \in [0.1, 0.5]$ with a step of 0.1. Such numbers have been chosen to mimic real-world settings.

3. FAST and PageRank

[7] was the earliest axiomatic review article of rank aggregation methods and discussed two classes of rank aggregation methods, namely ad hoc methods and distance or axiomatic-based methods. In this part, we are going to introduce two algorithms: PageRank (ad hoc method) and FAST(distance-based method), representing the ad hoc method and distance-based method respectively.

3.1 PageRank

The main idea of the ranking aggregation algorithm PageRank [15] is: given all base ranking lists, graph G = (V, E) can be constructed where each item is a node in the graph G. For each ranking where item m_i ranks higher than m_j, we have a directed edge (m_j, m_i) whose weight equals to difference in ranks. Then normalize the weights so that outgoing edges have a total weight of 1 for each node. The PageRank $PR(m_i)$ of an item m_i is

$$PR(m_i) = (1 - \alpha) \times p_i + \alpha \times \sum_{(m_j, m_i) \in E} \frac{PR(m_j) \times w(m_i, m_j)}{k_{m_j}^{out}} \tag{2}$$

where $k_{m_j}^{out}$ represents outdegree of node m_j, $k_{m_j}^{in}$ represents indegree of node m_j. The probability of randomly jumping to a node is proportional to the indegree of that node where $p_i = \frac{k_{m_i}^{in}}{\sum_{a_j \in V} k_{m_j}^{in}}$.

3.2 FAST

Kendall proposed τ_a and τ_b rank correlation coefficient and the former can only be used for FLR and the latter also can be used for WLR. Emond and Mason [3] proposed τ_x rank correlation coefficient based on τ_b for its problem of handling ties. In a list containing M items, we define score matrix A as an $M \times M$ matrix for any two item x and y, $A_{xy} = 1$ if x ranks higher than y or tie with y; $A_{xy} = -1$ if x ranks lower than y; $A_{xy} = 0$ if x and y are identical (e.g., element on diagonal line in A always equals to 0). It is clear that element 0 not on the diagonal line represents no comparison information between the corresponding two items. Given two list R_{μ_1} and R_{μ_2}, τ_x is defined as

$$\tau_x(R_{\mu_1}, R_{\mu_2}) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{ij}}{M(M-1)} \tag{3}$$

Emond and Mason also proved that the Kemeny-Snell distance metric and the τ_x rank correlation coefficient were equivalent representations of the unique measure that satisfied the Kemeny-Snell axioms. It is clear that any desirable properties of Kemeny-Snell distance metric are also desirable properties of the τ_x rank correlation coefficient. Given M items and N ranking lists $R_{\mu_1}, \ldots, R_{\mu_N}$, the problem is to find a ranking R_{μ^*} that maximizes the weighted average correlation with the input rankings or, equivalently, minimizes the weighted average Kemeny distance to the input rankings,

$$\max \frac{\sum_{k=1}^{N} w_k \tau_x\left(R_{\mu^*}, R_{\mu_k}\right)}{\sum_{k=1}^{N} w_k} \tag{4}$$

where w is weight vector specifying prior information about the importance or reliability of the input rankings. Here we assign all elements in w with 1 representing each input ranking is as important as others. Indicating as $\{r_{ij}^*\}$ and $\{r_{ij}\}^{(k)}$ the scoring matrices for R_{μ^*} and the kth ranking R_{μ_k}, k = 1 , …, N, the problem is:

$$\max \sum_{k=1}^{N} \left\{ \sum_{i=1}^{M} \sum_{j=1}^{M} r_{ij}^* r_{ij}{}^{(k)} \right\} = \max \sum_{i=1}^{M} \sum_{j=1}^{M} r_{ij}^* c_{ij} \tag{5}$$

where $c_{ij} = \sum_{k=1}^{n} r_{ij}^{(k)}$. The score matrix $\{c_{ij}\}$ was called Combined Input Matrix (CI) by Emond and Mason because it was the result of a summation of each input ranking. Defined in this way, it summarizes the ranking information in a single matrix. Based on the above function, Emond and Mason proposed a branch and bound algorithm based on which [16] proposed a new algorithm FAST with better computing efficiency and shorter time.

4. Performance Index

Performance index plays an important role in evaluating and comparing FAST and PageRank methods described above properly. A measure of the similarity between incomplete rankings should handle non-conjointness, weight high ranks more heavily than low, and be monotonic with increasing depth of evaluation. Such a measure with these features qualifies as an indefinite rank similarity measure. There are many indices in the field of rank aggregation, such as Kendall τ distance, Spearman's footrule, Hausdorff distance, Kendall τ correlation coefficient, etc. However, none of them meet all the features above. As far as we know, this paper is the first one to use rank-biased overlap in comparing rank aggregation methods, so we will pick one of the traditional indices to compare with it. In this section, two performance indices, namely generalized Kendall rank correlation coefficient [17] and rank-biased overlap [18] will be described in detail.

4.1 Generalized Kendall Rank Correlation Coefficient τ_g

The most commonly used index to evaluate the performance of algorithms is the Kendall τ distance in the field of rank aggregation comparison, which counts the number of pairs for which the order is different in two permutations. Slight modification on τ will be conducted to enable it suitable for all list types and in addition, we use its equivalent form, rank correlation coefficient, for better comparison with the rank-biased overlap. The generalized Kendall rank correlation coefficient τ_g is defined as:

$$\tau_g = \frac{n_c - n_d - n_u}{\binom{n}{2} - n_u} \tag{6}$$

where n_c and n_d are the numbers of concordant and discordant pairs respectively, n is the total number of unique items between the lists, and n_u is the number of unlabeled pairs. From the above

definition, $\tau_g \in [-1, 1]$ and it gives a higher penalty to disjoint partial lists than reverse partial lists. For the goal of better comparing with rank-biased overlap, after a simple linear transformation, we can have $\tau_g \in [0, 1]$.

4.2 Rank-biased Overlap

[18] proposed a new similarity measure, rank-biased overlap (RBO), which is based on a simple user model in which the user compares the overlap of two ranking at incrementally increasing depths. The author also proved that RBO meets all criteria for an indefinite rank similarity measure. It provides monotonicity by calculating, at a given depth of evaluation, a base score that is non-decreasing with additional evaluation, and a maximum score that is non-increasing. An extrapolated score can be calculated between these bounds if a point estimate is required. RBO can be applied to all kinds of list types and belongs to range from 0 to 1. We take user persistence p = 0.95. See [18] for more details.

5. Results and Discussion

This section mainly presents our experimental results illustrating the impact of data characteristics on the performance of rank aggregation methods. We use results under the combination of N = 15, M = 80, $r_t = 0.2$, $r_k = 0.8$ and $\Delta k = 0.2 * M$ without special statement.

5.1 Comparison between List Types with Various θ

Intuitively, the higher the degree of consensus θ behind the base rankers is, the better the data quality and the final aggregation effect will be. It can be seen from Fig. 1 that as θ increases gradually from basically no consensus θ = 0.001 to strong consensus θ = 0.7, the base rankers follow from a random distribution to a Gaussian-like distribution around σ_0, and both the τ_g and RBO become higher, especially for FLR and WLR types. When θ = 0.001, the initial set of the base rankers follows a uniform distribution representing low data quality at this time, and the aggregation effect is also poor with RBO < 0.4, which is analogous to the result from a dataset generated randomly. For all list types, the results are consistent in general, indicating that the degree of consensus θ or data quality directly determines the aggregation effect of various rank aggregation methods. It's worthy to note that when we process IR lists, it appears necessary to conduct pre-processing first before feeding them into algorithms.

Generally, there is no much difference between FAST and PageRank when θ is fixed in most cases, and the PageRank even performs little better than FAST in some experimental settings.

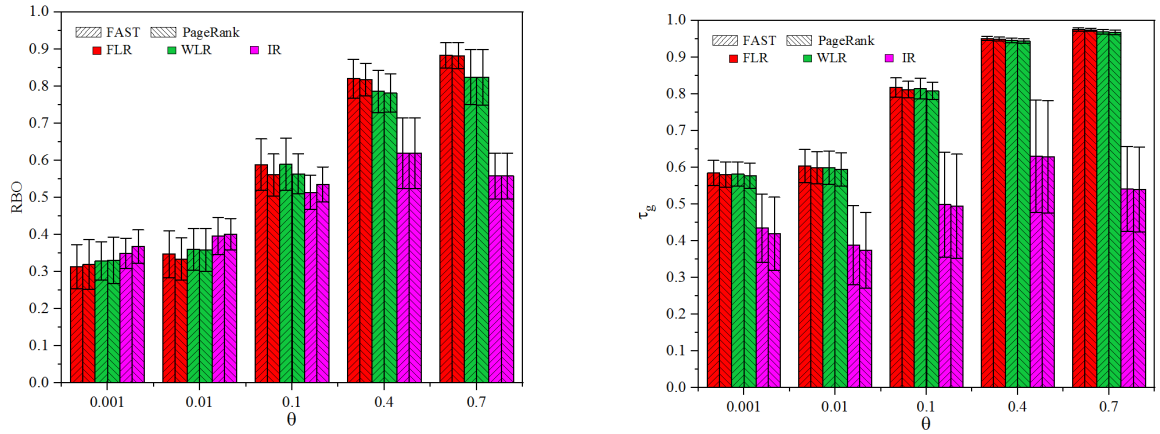

Fig. 1 Comparison between list types with various θ

5.2 Impact of Ties

The difference between the FLR list and the WLR list generated from our model is that the WLR list contains ties. This section explores the impact of the number of ties on the aggregation effect.

When the number of ties is small, the ties have little effect on the aggregated results. For methods that can handle ties, the result by breaking all the ties and randomly rearranging the corresponding items is slightly different from that by way of considering the cost of untying. Overall, results from the former way are a little worse than the latter. The same conclusion holds for similar kind of methods to perform the above process. On one hand, randomly rearranging the tied items changes the original ranking lists and becomes a different aggregation process. On the other hand, after the randomization operation, there must produce many candidate results from which one can not make a choice. This is the real reason why most methods need to be able to handle ties.

However, as ties become more and more gradually, will that make a difference for the aggregation results? In the data generation model described in section 2, the number of ties included in WLRs is uniformly distributed, so when N is fixed, the number of ties tends to increase uniformly just like the first plot in Fig. 2. As can be seen from the last two plots in Fig. 2, RBO and τ_g both are declining as the total number of ties continues to increase, indicating that the quality of the originally generated data has changed. Lists with $r_t = 0$ correspond to FLRs. When r_t is small, the ties make no big change and there is not much fluctuation at this time. As r_t gradually increases, the aggregation results of the algorithms become worse than that from FLR. In such cases, FAST performs better than PageRank.

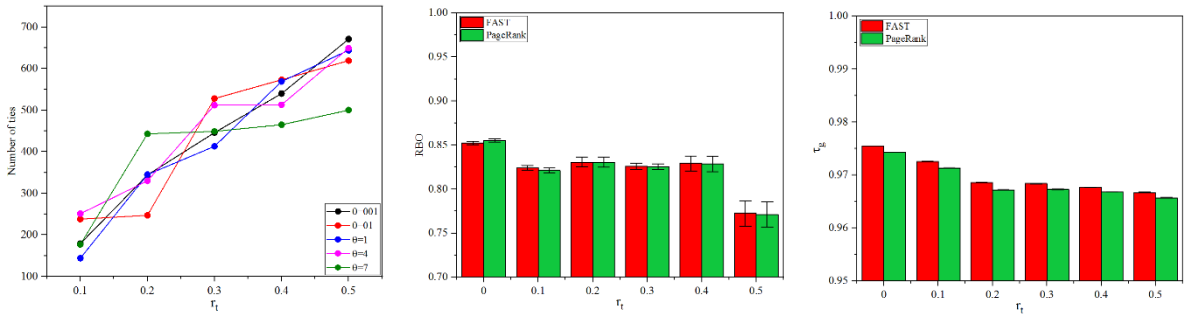

Fig. 2 Impact of ties

5.3 Impact of r_k

Sometimes, it is not necessary to consider all items in a ranking list, such as the case of results from search engines, and what we really care is the top few items. This part considers the impact of r_k on the performance of rank aggregation methods. Parameter r_k is proportional with the base length k of IR lists, which directly determines the lower and upper bound of the length of IR lists. The set of all IR lists under particular settings follows a uniform distribution as shown in section 2. For small θ in our experiments such as 0.001 or 0.01, IR lists follow a uniform distribution indicating the placement of items is generally random. At that moment, as r_k increases, the coverage rate always equals to 1. But if θ becomes larger such as 0.4 or 0.7, the coverage rate increases with r_k gradually to 1, as can be seen in the first plot in Fig. 3. Once θ fixed, we can find that both RBO and τ_g increases with r_k, this is because we will have more preference information as r_k becomes larger.

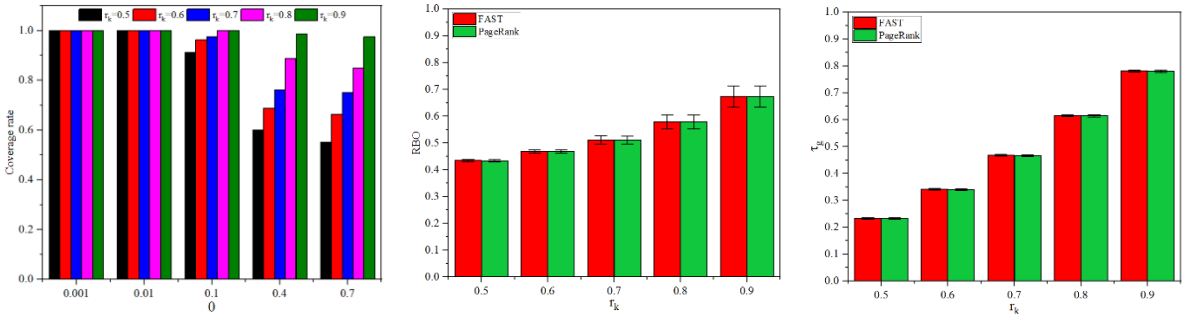

Fig. 3 Impact of r_k

5.4 Impact of N

Section 5.1 explores the impact of θ on the performance. Then, does the number of rankers N under fixed θ have a remarkable impact on the aggregation effect? To explore this problem, let θ = 0.7, M = 80, and the result is shown in Fig. 4. The results show that for all list types, the performance of FAST and PageRank show high stability with various parameter settings. It means that once the degree of consensus and transformation parameters chosen, the data quality is then determined regardless of how many rankers we have, and it will cause no big fluctuations.

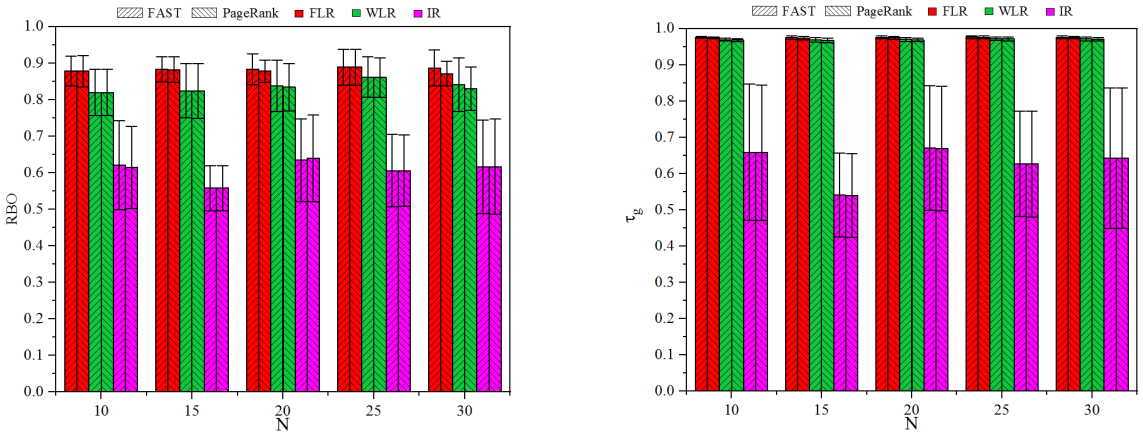

Fig. 4 Impact of N

6. Summary

This paper presents a general data generation mechanism based on Mallows model for all list types to produce synthetic controllable datasets, uses generalized Kendall rank correlation coefficient and rank-biased overlap to evaluate and compare the performance of two kinds of methods under different settings. Besides, we also consider the comparison between indices and the impact of data characteristics on the algorithms. The experimental results show that (1) data characteristics directly determine the performance of rank aggregation methods, such as the degree of consensus, number of ties, and other transformation strategies; (2) ad hoc methods sometimes performs better than distance-based methods with shorter time and better efficiency; (3) RBO is preferred than any other traditional indices, such as Kendall distance, Kendall rank correlation coefficient or their generalization. This paper may be helpful to researchers and decision-makers.

References

[1]. Snell J L, Kemeny J G. Mathematical Models in the Social Sciences[M]//Mathematical thinking in the social sciences/. Free Press, 1962.

[2]. Davenport A, Kalagnanam J. A computational study of the Kemeny rule for preference aggregation[C]//AAAI. 2004, 4: 697-702.

[3]. Emond E J, Mason D W. A new rank correlation coefficient with application to the consensus ranking problem[J]. Journal of Multi-Criteria Decision Analysis, 2002, 11(1): 17-28.

[4]. Lü L, MEDO M, YEUNG C H, et al. Recommender systems[J]. Physics reports, 2012, 519(1): 1-49.

[5]. Dwork C, Kumar R, Naor M, et al. Rank aggregation methods for the web[C]//Proceedings of the 10th international conference on World Wide Web. ACM, 2001: 613-622.

[6]. de Borda J C. Mémoire sur les élections au scrutiny [J]. 1781.

[7]. Cook W D. Distance-based and ad hoc consensus models in ordinal preference ranking[J]. European Journal of operational research, 2006, 172(2): 369-385.

[8]. Cook W D, Golany B, Penn M, et al. Creating a consensus ranking of proposals from reviewers' partial ordinal rankings[J]. Computers & Operations Research, 2007, 34(4): 954-965.

[9]. Ali A, Meilă M. Experiments with Kemeny ranking: What works when? [J]. Mathematical Social Sciences, 2012, 64(1): 28-40.

[10]. Xiao Y, Deng Y, Wu J, et al. Comparison of rank aggregation methods based on inherent ability[J]. Naval Research Logistics (NRL), 2017, 64(7): 556-565.

[11]. Schalekamp F, Zuylen A. Rank aggregation: Together we're strong[C]//2009 Proceedings of the Eleventh Workshop on Algorithm Engineering and Experiments (ALENEX). Society for Industrial and Applied Mathematics, 2009: 38-51.

[12]. Brancotte B, Yang B, Blin G, et al. Rank aggregation with ties: Experiments and analysis[J]. Proceedings of the VLDB Endowment, 2015, 8(11): 1202-1213.

[13]. Fagin R, Kumar R, Mahdian M, et al. Comparing and aggregating rankings with ties[C]//Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. ACM, 2004: 47-58.

[14]. Critchlow D E, Fligner M A, Verducci J S. Probability models on rankings[J]. Journal of mathematical psychology, 1991, 35(3): 294-318.

[15]. Adali S, Hill B, Magdon-Ismail M. Information vs. robustness in rank aggregation: Models, algorithms and a statistical framework for evaluation[J]. Journal of Digital Information Management, 2007, 5(5): 292.

[16]. Amodio S, D'Ambrosio A, Siciliano R. Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the Kemeny axiomatic approach[J]. European Journal of Operational Research, 2016, 249(2): 667-676.

[17]. Langville A N, Meyer C D. Who's# 1? : the science of rating and ranking[M]. Princeton University Press, 2012.

[18]. Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings[J]. ACM Transactions on Information Systems (TOIS), 2010, 28(4): 20.