Quantifying consensus of rankings based on q-support patterns

Zhengui Xue, Zhiwei Lin*, Hui Wang, Sally McClean

School of Computing, University of Ulster, United Kingdom

Abstract

Rankings, representing preferences over a set of candidates, are widely used in many applications, e.g., group decision making and information retrieval. Rankings may be obtained by different agents (humans or systems). It is often necessary to evaluate consensus of obtained rankings from multiple agents, as a measure of consensus provides insights into the rankings. Moreover, a consensus measure could provide a quantitative basis for comparing groups and for improving a ranking system. Existing studies on consensus measurement are insufficient, since they did not evaluate consensus among most rankings or consensus with respect to specific preference patterns. In this paper, a novel consensus quantifying approach, without the use of correlation or distance functions as in existing studies of consensus, is proposed based on the concept of q-support patterns, which represent the commonality embedded in a set of rankings. A pattern is regarded as a q-support pattern if it is included by at least q rankings in the ranking set. A method for detecting outliers in a set of rankings is naturally derived from the proposed consensus quantifying approach. Experimental studies are conducted to demonstrate the effectiveness of the proposed approach.

Keywords: Rankings, consensus, support patterns, outlier detection

*Corresponding author

Email addresses: zhenguixue@gmail.com (Zhengui Xue), z.lin@ulster.ac.uk (Zhiwei Lin), h.wang@ulster.ac.uk (Hui Wang), si.mcclean@ulster.ac.uk (Sally McClean)

1. Introduction

Extensive studies have been carried out in social science to measure group cohesion, in order to gain insight into the factors affecting group cohesion and further promote higher group consistency (see, e.g., [21, 7, 35, 10]). In artificial intelligence, rankings have been widely used to represent the preferences of agents (humans or systems) over a set of candidates in many information systems, such as group decision making [27, 34, 43] and information retrieval [22, 29, 33]. It is important to evaluate the degree to which the rankings obtained by different agents agree, as it would help to understand the obtained rankings. Quantifying the *consensus* of the obtained rankings can provide an accurate evaluation about the overall agreement. It is also a quantitative indicator for comparing consensus between groups (e.g., two sets of rankings) [4] or for further improving the ranking systems. For example, in group decision making, if the consensus score is extremely low, it is necessary for the experts to adjust their rankings in order to reach an agreement [27]. However, to the best of our knowledge, there are only a few existing studies [1, 2, 4, 13, 16, 17] on consensus evaluation for a set of rankings.

In the literature, rank *correlation* and *distance* functions, such as Kendall's τ [26] and Spearman's ρ [37], are used to measure the correlation and disagreement of two rankings. Kendall's τ measures the correlation of two rankings by considering their concordant and discordant pairs, and Spearman's ρ evaluates the rank correlation by taking into account the positions of the items in two rankings. The Kemeny distance [25] is extended to measure pairwise disagreements in two rankings. For a set with more than two rankings, the related concepts are *consensus* and *diversity* of rankings. Consensus is also used interchangeably for *cohesiveness* [2]. Existing approaches measure consensus of rankings by considering the similarity of preferences in a group based on rank correlation functions. One typical approach as discussed in [2] is to calculate the similarity for each pair of rankings based on correlation functions, such as Kendall's τ and Spearman's ρ, and then aggregate the obtained results. *Diversity* and *consensus* are considered as two opposite concepts about rankings in social choice theory [24]. Research was carried out to measure the diversity of a ranking set based on distance

functions (see [16]). These existing studies are not sufficient in evaluating the overall consensus of a ranking set. It is difficult to use the rank correlation or distance functions based approaches to completely quantify the level of consensus for a set of rankings. As pointed out in [12], the pairwise comparison reflects the degree of commonality in two rankings, and consequently the aggregated result of the pairwise comparisons is not informative enough to tell the degree to which the ranking set agrees. In reality, it is often the case that certain preference patterns are embedded in most of the rankings obtained for a task. The existing work cannot tell the degree to which preferences over candidates are shared by the majority of the rankings. In addition, they did not provide a solution to identifying the majority of rankings in order to filter irrelevant results in the ranking set, which could play an important role in modern information systems. For instance, in query expansion [6], it is reasonable to expand a query for 'film' to its relevant query 'movie' in the context of entertainment, but not in the context of 'a thin coat or layer'. It is impossible to manually check if the expansions from the source query are consistent as there is no ground truth available and moreover the meaning of the queries may evolve from time to time (e.g, 'apple' in fruit context to the context of cooperation). Therefore, using the rankings obtained from the expansions to understand the extent to which the query expansions provide high level of consistency is key to provide good search results.

This paper studies the consensus degree of a ranking set from a different perspective to provide a full picture on the degree to which a set of rankings mutually agree. A novel framework is proposed to analyze consensus of rankings by considering the common patterns embedded in a ranking set. A new concept of q-support patterns is introduced to represent how common patterns are embedded in rankings, by which the preferences of a group over candidates can be expressed at a subtle and fine-grained level. A pattern is regarded as a q-support pattern if it is included by at least q rankings in the ranking set. Thus, a q-support pattern represents the partial coverage of the pattern by rankings, where the integer q can be specified as needed when a ranking system is evaluated. The consensus degree of rankings is quantified based on q-support patterns. Compared with the existing work based on correlation or distance functions, this new approach gives a finer characterization and quantification of the commonality embedded in the rankings.

The contributions of this paper are: (1) a new representation of the commonality within a set of rankings, q-support pattern, is proposed; (2) a new framework (non-distance and non-correlation) for quantifying consensus with q-support patterns is introduced; (3) an efficient algorithm is developed to calculate consensus scores and characterize the set of q-support patterns; (4) consensus scores are defined for each ranking to reflect its relationship with the other rankings, which can be used to detect outliers in a ranking set; (5) extensive experiments have been conducted to show the effectiveness and usefulness of the proposed approach.

The rest of the paper is organized as follows. In Section 2, related work on the comparison of two rankings and the measure of consensus and diversity of a ranking set is reviewed. In Section 3, the q-support pattern of rankings is formulated and consensus scores are defined based on it. An algorithm is then introduced to calculate ranking consensus. In Section 4, weighted consensus scores are defined. In Section 5, an outlier detection method is developed. Section 6 gives experimental studies to evaluate the proposed approach. Section 7 concludes this paper.

2. Related work

Rank correlation and distance functions. Historically developed by Maurice Kendall in 1938 [26], Kendall's τ measures the correlation between two rankings by considering the numbers of pairwise items ranked in same orders and in opposite orders. Suppose that we consider rankings over candidates $\{\sigma_1, \sigma_2, \cdots, \sigma_n\}$. A ranking is an ordered list in which items in higher positions are more preferred than items in lower positions. Let $\pi(\cdot,\cdot)$ be the position function. The function $\pi(\sigma_i, \mathbf{r}_l)$ returns the position of item σ_i in ranking $\mathbf{r}_l$. Kendall's τ for two rankings $\mathbf{r}_l$ and $\mathbf{r}_z$ is

$$\tau(\mathbf{r}_l, \mathbf{r}_z) = \frac{\sum\limits_{\substack{i,j\in\{1,\cdots,n\}\\ i<j}} \text{sgn}(\pi(\sigma_i, \mathbf{r}_l) - \pi(\sigma_j, \mathbf{r}_l))\text{sgn}(\pi(\sigma_i, \mathbf{r}_z) - \pi(\sigma_j, \mathbf{r}_z))}{n(n-1)/2}.$$

This coefficient is in the range $-1 \leq \tau(\mathbf{r}_l, \mathbf{r}_z) \leq 1$, where value 1 corresponds to the case that the two rankings are in the same order and value -1 indicates that one ranking is in the reverse order of the other.

Spearman's ρ proposed by Charles Spearman in 1904 [37] is defined based on the positions of each item in two rankings as follows

$$\rho(\mathbf{r}_l, \mathbf{r}_z) = \frac{\sum_{i=1}^{n}(\pi(\sigma_i, \mathbf{r}_l) - \bar{\pi}_l)(\pi(\sigma_i, \mathbf{r}_z) - \bar{\pi}_z)}{\sqrt{\sum_{i=1}^{n}(\pi(\sigma_i, \mathbf{r}_l) - \bar{\pi}_l)^2 \sum_{i=1}^{n}(\pi(\sigma_i, \mathbf{r}_z) - \bar{\pi}_z)^2}},$$

where $\bar{\pi}_l = \frac{1}{n}\sum_{i=1}^{n}\pi(\sigma_i, \mathbf{r}_l)$ and $\bar{\pi}_z = \frac{1}{n}\sum_{i=1}^{n}\pi(\sigma_i, \mathbf{r}_z)$. Similarly, this coefficient satisfies $-1 \leq \rho(\mathbf{r}_l, \mathbf{r}_z) \leq 1$.

These rank correlation functions do not take into account the varying relevance of ranked items in different positions. They are not suitable for evaluating the rankings where items at the top of a ranking are much more important than those at the bottom [15]. Further studies on weighted rank correlation were carried out extensively based on these two functions [8, 23, 28, 36, 39, 41, 42]. More reasonable variants of rank correlation functions were also proposed in the literature [14, 19, 20, 38].

Distance metrics have been used to analyze ranking data. One of the most widely used distance functions to measure rankings is the Kemeny distance [25]. It is defined as the sum of pairs where the ranking preferences disagree. One can refer to [3, 31, 11] for more information about the commonly used distance metrics.

Measuring consensus and diversity of rankings. For a ranking set with the number of rankings greater than two, work [4] is known as the first study to define a consensus measure as a function mapping linear orders (i.e, rankings without ties) to a number between 0 and 1. Kendall's coefficient of concordance was introduced in [4] as a measure of consensus of a ranking set. Given a set of rankings $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$ over candidates $\{\sigma_1, \sigma_2, \cdots, \sigma_n\}$, the total positions of the candidates in all rankings need to be calculated first, which are $\sum_{l=1}^{N}\pi(\sigma_i, \mathbf{r}_l), i = 1, \cdots, n$. Kendall's coefficient of concordance is defined based on the deviations of the total positions from their mean as

$$W = \frac{12}{N^2(n^3 - n)}\sum_{i=1}^{n}\left(\sum_{l=1}^{N}\pi(\sigma_i, \mathbf{r}_l) - \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{N}\pi(\sigma_i, \mathbf{r}_l)\right)^2,$$

where the term $\frac{12}{N^2(n^3-n)}$ is for normalization.

García-Lapresta and Pérez-Román [16] extended the work [4] by considering weak

orders (i.e., ranking with ties). A measure based on a weighted Kemeny distance was introduced. In [2], it was discussed that one prominent approach of constructing a consensus or diversity measure is to make pairwise comparisons of the rankings with a rank correlation or distance function, such as the functions introduced in the above section, and then aggregate the comparison results. Thus, two key issues with this approach are the choice of a proper pairwise comparison metric and the utilization of an aggregation method. Kendall's τ was used to compare the similarity of each pair of rankings in [2], and the consensus measure of a ranking set was constructed by taking the average of the comparison results. Studies with more reasonable similarity or distance metrics were carried out in [1, 17, 18, 13]. Karpov [24] considered to aggregate the comparison results with a geometric mean aggregator.

Although these studies discussed different aspects of consensus measures, they are still inefficient in the assessment of overall consensus of a ranking set. In information systems, it is often the case that certain preference patterns are embedded in most of the rankings. The existing studies based on rank correlation and distance functions did not provide a full picture about this kind of common patterns. They cannot quantify the degree to which preferences over candidates are shared by the majority of the rankings. To solve this problem, this paper proposes a concept of q-support patterns to represent the commonality in a ranking set and the consensus is quantified based on the q-support patterns.

3. Quantifying consensus with q-support patterns

This section first defines the q-support patterns and the consensus scores of a ranking set. Then, an algorithm is presented to calculate the consensus scores by utilizing matrices to represent the q-support patterns.

3.1. q-support patterns

Let $\mathcal{C} = \{\sigma_1, \sigma_2, \cdots, \sigma_n\}$ be a set of n candidates to be ranked. A ranking $\mathbf{r}_l = (r_{l_1}, r_{l_2}, \cdots, r_{l_m})$ is an ordered list in which item $r_{l_i} \in \mathcal{C}$ is more preferred than item $r_{l_j} \in \mathcal{C}$ for $i < j$. Given two items σ_x and $\sigma_y \in \mathcal{C}$, if there exists $i \leq j$ such that

$r_{l_i} = \sigma_x$ and $r_{l_j} = \sigma_y$, we write $\sigma_x\sigma_y \sqsubset \mathbf{r}_l$; otherwise $\sigma_x\sigma_y \not\sqsubset \mathbf{r}_l$. Specially, if $\sigma_x = \sigma_y$, $\sigma_x\sigma_x \sqsubset \mathbf{r}_l$ simply means that item σ_x is included in ranking $\mathbf{r}_l$, also written as $\sigma_x \sqsubset \mathbf{r}_l$.

It is usually the case that most of the rankings obtained for a task share certain commonality. Suppose that there is a set of rankings $\mathcal{R} = \{\mathbf{r}_1 = (a, b, c, d, e, f),\ \mathbf{r}_2 = (b, a, c, d, e, f),\ \mathbf{r}_3 = (a, b, c, e, d, f),\ \mathbf{r}_4 = (c, b, d, e, f, g)\}$. It can be seen that item a and the pairwise items bc are common patterns for most of the rankings, but not for all the rankings in $\mathcal{R}$ (e.g., $bc \sqsubset \mathbf{r}_1, bc \sqsubset \mathbf{r}_2, bc \sqsubset \mathbf{r}_3$, but $bc \not\sqsubset \mathbf{r}_4$). These patterns, partially included in a set of rankings, show the extent to which the rankings agree. Therefore, it is necessary to consider these patterns to understand the consensus level in a set of rankings. As such, we define the following q-support patterns for a ranking set.

Definition 1 (q-support patterns)**.** *Consider a set of N rankings $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$ over candidate set $\mathcal{C} = \{\sigma_1, \sigma_2, \cdots, \sigma_n\}$. For σ_x and $\sigma_y \in \mathcal{C}$, we have the following subset $\mathcal{R}'(\sigma_x, \sigma_y) \subseteq \mathcal{R}$*

$$\mathcal{R}'(\sigma_x, \sigma_y) = \{\mathbf{r}_z | \sigma_x\sigma_y \sqsubset \mathbf{r}_z,\ \mathbf{r}_z \in \mathcal{R}\}. \tag{1}$$

Let $q \in (0, N]$ be an integer. The pattern $\sigma_x\sigma_y$ is a q-support pattern of $\mathcal{R}$, denoted by $\sigma_x\sigma_y \overset{q}{\sqsubset} \mathcal{R}$, if the size of $\mathcal{R}'(\sigma_x, \sigma_y)$ satisfies $|\mathcal{R}'(\sigma_x, \sigma_y)| \geq q$; otherwise $\sigma_x\sigma_y \overset{q}{\not\sqsubset} \mathcal{R}$. If $\sigma_x = \sigma_y$, $\sigma_x\sigma_x \overset{q}{\sqsubset} \mathcal{R}$ indicates that item σ_x is a single q-support item of $\mathcal{R}$, also written as $\sigma_x \overset{q}{\sqsubset} \mathcal{R}$.

The notation $\sigma_x\sigma_y \overset{q}{\sqsubset} \mathcal{R}$ means that $\sigma_x\sigma_y$ occurs in at least q rankings in $\mathcal{R}$. We use $\mathbf{S}_1(q)$ and $\mathbf{S}_2(q)$ to respectively denote the set of the single q-support items and the set of the pairwise q-support patterns, i.e.,

$$\mathbf{S}_1(q) = \left\{\sigma_x | \sigma_x \overset{q}{\sqsubset} \mathcal{R},\ \sigma_x \in \mathcal{C}\right\} \tag{2}$$

$$\mathbf{S}_2(q) = \left\{\sigma_x\sigma_y | \sigma_x\sigma_y \overset{q}{\sqsubset} \mathcal{R},\ \sigma_x \neq \sigma_y, \sigma_x \in \mathcal{C}, \sigma_y \in \mathcal{C}\right\}. \tag{3}$$

The set $\mathbf{S}_1(q)$ is important in the evaluation of incomplete rankings, where not all the candidates under consideration are ranked in the rankings. It gives the items with more preferences among the candidates, which are ranked in at least q rankings. The set $\mathbf{S}_2(q)$ collects the preference orders embedded in at least q rankings.

3.2. Consensus scores

The q-support patterns describe how common patterns are embedded in rankings. This section first defines individual consensus scores for a ranking $\mathbf{r}_l \in \mathcal{R}$ based on the q-support patterns. Then, the overall consensus scores are introduced for the ranking set $\mathcal{R}$. The relative consensus degree that a ranking $\mathbf{r}_l$ shares with the others can be revealed by the individual and the overall consensus scores. In Section 5, it shows that this information can be used in the detection of an outlier from a ranking set.

The following individual consensus scores are defined for a ranking $\mathbf{r}_l$.

Definition 2 (Individual consensus scores)**.** *For a ranking $\mathbf{r}_l = (r_{l_1}, r_{l_2}, \cdots, r_{l_m}) \in \mathcal{R}$, the sets of the single q-support items and the pairwise q-support patterns are defined as*

$$\mathbf{S}_1^{\mathbf{r}_l}(q) = \left\{ r_{l_i} | r_{l_i} \overset{q}{\sqsubset} \mathcal{R},\ i \in \{1, 2, \cdots, m\} \right\} \tag{4}$$

$$\mathbf{S}_2^{\mathbf{r}_l}(q) = \left\{ r_{l_i} r_{l_j} | r_{l_i} r_{l_j} \overset{q}{\sqsubset} \mathcal{R},\ i, j \in \{1, 2, \cdots, m\}, i < j \right\}. \tag{5}$$

The individual consensus scores of $\mathbf{r}_l$ are

$$\kappa_1^{\mathbf{r}_l}(q) = \frac{1}{N_1^{\mathbf{r}_l}} |\mathbf{S}_1^{\mathbf{r}_l}(q)| \tag{6}$$

$$\kappa_2^{\mathbf{r}_l}(q) = \frac{1}{N_2^{\mathbf{r}_l}} |\mathbf{S}_2^{\mathbf{r}_l}(q)|, \tag{7}$$

where $N_1^{\mathbf{r}_l} = m$ and $N_2^{\mathbf{r}_l} = \frac{m(m-1)}{2}$ respectively represent the number of the ranked items and the number of the pairwise patterns of $\mathbf{r}_l$.

Definition 3 (Overall consensus scores)**.** *For a ranking set $\mathcal{R}$ with the individual consensus scores defined as (6) and (7), the overall consensus scores of $\mathcal{R}$ are*

$$\bar{\kappa}_1(q) = \frac{1}{N} \sum_{l=1}^{N} \kappa_1^{\mathbf{r}_l}(q) \tag{8}$$

$$\bar{\kappa}_2(q) = \frac{1}{N} \sum_{l=1}^{N} \kappa_2^{\mathbf{r}_l}(q). \tag{9}$$

The individual consensus scores measure the proportions of the preference patterns of $\mathbf{r}_l$ embedded in at least q rankings, where $\kappa_1^{\mathbf{r}_l}(q)$ measures consensus in terms of single q-support items and $\kappa_2^{\mathbf{r}_l}(q)$ measures consensus in terms of pairwise q-support

patterns. The overall consensus scores give the average proportions and they are used to evaluate the consensus degree of a whole ranking set. Note that a q-support pattern depicts the commonality embedded in at lease q rankings in a ranking set. The choice of q in the consensus evaluation depends on the specific need in the evaluation of ranking data. For example, many information systems may expect that ranked patterns are supported by at least half of the experts, and the value of q can be set to $\lceil \frac{N}{2} \rceil$ for this case. In addition, by studying the consensus degree based on different values of q, a more comprehensive understanding about the ranking set can be obtained, as different values of q reflect the extents of different partial coverage of the patterns embedded in rankings.

The consensus scores have the following property.

Property 1. The overall consensus scores satisfy

$$0 \leq \bar{\kappa}_1(q) \leq 1 \tag{10}$$

$$0 \leq \bar{\kappa}_2(q) \leq 1. \tag{11}$$

The score $\bar{\kappa}_1(q) = 0$ if and only if arbitrary q rankings in $\mathcal{R}$ share no common item, and $\bar{\kappa}_1(q) = 1$ if and only if every ranked item of all the rankings is shared by at least q rankings. Similarly, $\bar{\kappa}_2(q) = 0$ if and only if arbitrary q rankings in $\mathcal{R}$ share no common pairwise pattern, and $\bar{\kappa}_2(q) = 1$ if and only if every pairwise preference pattern of all the rankings is embedded in at least q rankings.

3.3. An efficient algorithm for quantifying consensus

In this section, a matrix representation is introduced to represent the q-support patterns, as shown in Theorem 1, which implies an algorithm for calculating the consensus scores.

Theorem 1. *Consider a set of N rankings $\mathcal{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$ over candidates $\mathcal{C} = \{\sigma_1, \sigma_2, \cdots, \sigma_n\}$. For a ranking $\mathbf{r}_l = (r_{l_1}, r_{l_2}, \cdots, r_{l_m}) \in \mathcal{R}$ and $\forall \mathbf{r}_z = (r_{z_1}, r_{z_2}, \cdots, r_{z_u}) \in \mathcal{R}$, with the position function*

$$\pi\left(r_{l_i}, \mathbf{r}_z\right) = \begin{cases} 0, & \text{if } r_{l_i} \not\sqsubset \mathbf{r}_z \\ p, & \text{if } r_{l_i} = r_{z_p} \end{cases} \tag{12}$$

and the Heaviside function

$$H(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

we define

$$f(r_{l_i}, r_{l_j}) = \begin{cases} \sum_{z=1}^{N} H\left(\pi(r_{l_i}, \mathbf{r}_z)\right), & \text{if } i = j \\ \sum_{z=1}^{N} H\left(\pi(r_{l_j}, \mathbf{r}_z) - \pi(r_{l_i}, \mathbf{r}_z)\right) H\left(\pi(r_{l_i}, \mathbf{r}_z)\right), & \text{otherwise} \end{cases} \tag{14}$$

and matrix $\mathbf{A}^{\mathbf{r}_l} = (A^{\mathbf{r}_l}[j,i]) \in \mathbb{R}^{m \times m}$ *as*

$$A^{\mathbf{r}_l}[j,i] = \begin{cases} 1, & \text{if } i \leq j \text{ and } f(r_{l_i}, r_{l_j}) \geq q \\ 0, & \text{otherwise}. \end{cases} \tag{15}$$

Then, we have

$$\kappa_1^{\mathbf{r}_l}(q) = \frac{1}{N_1^{\mathbf{r}_l}} tr\left(\mathbf{A}^{\mathbf{r}_l}\right) \tag{16}$$

$$\kappa_2^{\mathbf{r}_l}(q) = \frac{1}{N_2^{\mathbf{r}_l}} \left(\mathbf{e}^T \mathbf{A}^{\mathbf{r}_l} \mathbf{e} - tr\left(\mathbf{A}^{\mathbf{r}_l}\right)\right), \tag{17}$$

where $\mathbf{e} = [1, 1, \cdots, 1]^T$ *is an* m*-row vector of all ones.*

Proof. By (12), it can be known that $\pi(r_{l_i}, \mathbf{r}_z)$ gives the position of item r_{l_i} in $\mathbf{r}_z$. From the definition of $f(r_{l_i}, r_{l_j})$, it can be seen that $f(r_{l_i}, r_{l_j})$ counts the number of rankings $\forall \mathbf{r}_z \in \mathcal{R}$ satisfying $r_{l_i} r_{l_j} \sqsubset \mathbf{r}_z$. Thus, the entry $A^{\mathbf{r}_l}[j,i] = 1$ represents $r_{l_i} r_{l_j} \overset{q}{\sqsubset} \mathcal{R}$. Moreover, note that $\mathbf{e}^T \mathbf{A}^{\mathbf{r}_l} \mathbf{e}$ gives the sum of the all entries in matrix $\mathbf{A}^{\mathbf{r}_l}$. Therefore, the result of (16) and (17) can be further obtained based on Definition 2. □

The matrix $\mathbf{A}^{\mathbf{r}_l}$ provides a proper representation of the q-support patterns in $\mathbf{r}_l$. This representation can further facilitate the analysis of the commonality that individual rankings share with the others. Based on Theorem 1, we develop Algorithm 1 to calculate the consensus scores and characterize the q-support patterns more efficiently. In Algorithm 1, when a ranking $\mathbf{r}_l$ is considered, for a pattern $r_{l_i} r_{l_j}$ embedded in the ranking, there is no need to judge if the pattern is a q-support pattern by checking all the rankings in some cases. Suppose $q = \lceil \frac{2N}{3} \rceil$, which means that we consider

$r_{l_i}r_{l_j}$ as a common pattern if it is contained by at least two third of the rankings. For $\mathbf{r}_l$, if $r_{l_i}r_{l_j}$ is a $\lceil\frac{2N}{3}\rceil$-support pattern, it must be included by one of the rankings $\mathbf{r}_x \in \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{\lfloor\frac{N}{3}\rfloor+1}\}$. If $r_{l_i}r_{l_j}$ is not included by one of the first $\lfloor\frac{N}{3}\rfloor+1$ rankings of the ranking set, $r_{l_i}r_{l_j}$ cannot be a q-support pattern and $A^{\mathbf{r}_l}[j,i]$ should be zero. Thus, we do not need to calculate $A^{\mathbf{r}_l}[j,i]$ by always checking all the rankings. Line 7 in Algorithm 1 checks if $r_{l_i}r_{l_j}$ of $\mathbf{r}_l$ is included by a ranking $\mathbf{r}_x$ for which matrix $\mathbf{A}^{\mathbf{r}_x}$ has already been constructed. If the number of the rankings whose corresponding matrix is not constructed is greater than q, we look for $\mathbf{r}_x$ in the previously considered rankings $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{l-1}\}$; Otherwise if the number of rankings not yet considered is less than q, we only need to check if there is an $\mathbf{r}_x$ in the first $N-q+1$ rankings. As shown on Lines 8 and 9, if $r_{l_i}r_{l_j}$ has been considered in a constructed matrix for $\mathbf{r}_x$, it is not necessary to recalculate the corresponding entry of the current matrix $\mathbf{A}^{\mathbf{r}_l}$ and the entry is equal to that of $\mathbf{A}^{\mathbf{r}_x}$ corresponding to the pattern. Otherwise, as on Line 10, only when the number of the rankings $\{\mathbf{r}_l, \mathbf{r}_{l+1}, \ldots, \mathbf{r}_N\}$ is no less than q, pattern $r_{l_i}r_{l_j}$ has the possibility to be a q-support pattern and we need to check the remaining rankings to see if the pattern is a q-support ranking. In this way, the computation cost can be significantly reduced. From Lines 11 to 16, $f(r_{l_i}, r_{l_j})$ accumulates the number of rankings containing $r_{l_i}r_{l_j}$. To further improve the computation efficiency, the sum of $f(r_{l_i}, r_{l_j})$ and the number of the remaining rankings not yet considered is checked during the accumulation process. If it is less than q, then $r_{l_i}r_{l_j}$ has no chance to be a q-support pattern and there is no need to check if the remaining rankings contain $r_{l_i}r_{l_j}$. In Algorithm 1, for the case that $r_{l_i}r_{l_j}$ has no chance to be a q-support pattern, $A^{\mathbf{r}_l}[j,i]$ keeps the initialized value, i.e., zero.

The following example shows how the matrix representation can be used to evaluate the ranking consensus.

Example 1. *Consider a set of rankings* $\mathcal{R} = \{\mathbf{r}_1 = (a,b,c,d,e,f), \mathbf{r}_2 = (b,c,d,e,f,a), \mathbf{r}_3 = (b,d,a,g,h,f), \mathbf{r}_4 = (b,a,c,d,f,e)\}$ *over candidates* $\{a,b,c,d,e,f,g,h\}$*, and let* $q = 3$.

We have

$$\mathbf{A}^{\mathbf{r}_1}=\begin{array}{c} \\ a\\ b\\ c\\ d\\ e\\ f\end{array}\!\!\begin{array}{c}\begin{array}{cccccc} a & b & c & d & e & f\end{array}\\ \begin{pmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&1&1&0&0&0\\0&1&1&1&0&0\\0&1&1&1&1&0\\1&1&1&1&0&1\end{pmatrix}\end{array},\ \mathbf{A}^{\mathbf{r}_2}=\begin{array}{c} \\ b\\ c\\ d\\ e\\ f\\ a\end{array}\!\!\begin{array}{c}\begin{array}{cccccc} b & c & d & e & f & a\end{array}\\ \begin{pmatrix}1&0&0&0&0&0\\1&1&0&0&0&0\\1&1&1&0&0&0\\1&1&1&1&0&0\\1&1&1&0&1&0\\1&0&0&0&0&1\end{pmatrix}\end{array}$$

$$\mathbf{A}^{\mathbf{r}_3}=\begin{array}{c} \\ b\\ d\\ a\\ g\\ h\\ f\end{array}\!\!\begin{array}{c}\begin{array}{cccccc} b & d & a & g & h & f\end{array}\\ \begin{pmatrix}1&0&0&0&0&0\\1&1&0&0&0&0\\1&0&1&0&0&0\\0&0&0&0&0&0\\0&0&0&0&0&0\\1&1&1&0&0&1\end{pmatrix}\end{array},\ \mathbf{A}^{\mathbf{r}_4}=\begin{array}{c} \\ b\\ a\\ c\\ d\\ f\\ e\end{array}\!\!\begin{array}{c}\begin{array}{cccccc} b & a & c & d & f & e\end{array}\\ \begin{pmatrix}1&0&0&0&0&0\\1&1&0&0&0&0\\1&0&1&0&0&0\\1&0&1&1&0&0\\1&1&1&1&1&0\\1&0&1&1&0&1\end{pmatrix}\end{array}.$$

By (16) and (17), the following result can be obtained

l	1	2	3	4
$\kappa_1^{\mathbf{r}_l}(3)$	*1.00*	*1.00*	*0.67*	*1.00*
$\kappa_2^{\mathbf{r}_l}(3)$	*0.67*	*0.67*	*0.33*	*0.73*

The overall consensus scores are

$$\bar{\kappa}_1(3)=0.92,\ \bar{\kappa}_2(3)=0.60.$$

Since $A^{\mathbf{r}_l}[j,i]$ represents if $r_{l_i}r_{l_j}$ is a q-support pattern, it can be known $\mathbf{S}_1^{\mathbf{r}_1}(3)=\{a,b,c,d,e,f\}$, $\mathbf{S}_2^{\mathbf{r}_1}(3)=\{af,bc,bd,be,bf,cd,ce,cf,de,df\}$, $\mathbf{S}_1^{\mathbf{r}_2}(3)=\{b,c,d,e,f,a\}$, $\mathbf{S}_2^{\mathbf{r}_2}(3)=\{bc,bd,be,bf,ba,cd,ce,cf,de,df\}$, $\mathbf{S}_1^{\mathbf{r}_3}(3)=\{b,d,a,f\}$, $\mathbf{S}_2^{\mathbf{r}_3}(3)=\{bd,ba,bf,df,af\}$, $\mathbf{S}_1^{\mathbf{r}_4}(3)=\{b,a,c,d,f,e\}$, $\mathbf{S}_2^{\mathbf{r}_4}(3)=\{ba,bc,bd,bf,be,af,cd,cf,ce,df,de\}$. *Furthermore, the sets of the q-support patterns of the whole ranking set are* $\mathbf{S}_1(3)=\mathbf{S}_1^{\mathbf{r}_1}(3)\cup\mathbf{S}_1^{\mathbf{r}_2}(3)\cup\mathbf{S}_1^{\mathbf{r}_3}(3)\cup\mathbf{S}_1^{\mathbf{r}_4}(3)=\{a,b,c,d,e,f\}$, $\mathbf{S}_2(3)=\mathbf{S}_2^{\mathbf{r}_1}(3)\cup\mathbf{S}_2^{\mathbf{r}_2}(3)\cup\mathbf{S}_3^{\mathbf{r}_3}(3)\cup\mathbf{S}_4^{\mathbf{r}_4}(3)=\{af,ba,bc,bd,be,bf,cd,ce,cf,de,df\}$.

Algorithm 1: Quantifying consensus with matrix representation

Data: A set of rankings $\mathcal{R}$, the value of q

Result: $\kappa_1^{\mathbf{r}_l}(q), \kappa_2^{\mathbf{r}_l}(q), l = 1, 2, \cdots, N$; $\mathbf{S}_1^{\mathbf{r}_l}(q)$, $\mathbf{S}_2^{\mathbf{r}_l}(q)$; $\bar{\kappa}_1(q), \bar{\kappa}_2(q)$; $\mathbf{S}_1(q), \mathbf{S}_2(q)$

Initialize $\mathbf{A}^{\mathbf{r}_l}, l = 1, 2, \cdots, N$ with zero matrices
for $l = 1$ *to* N **do**
  $m \leftarrow$ Length of $\mathbf{r}_l$
  for $i = 1$ *to* m **do**
    for $j = i$ *to* m **do**
      $f(r_{l_i}, r_{l_j}) = 0$
      if $(l > 1, N - l + 1 \geq q, \exists x \in [1, l-1])$ *or* $(l > 1, N - l + 1 < q, \exists x \in [1, N - q + 1])$ *such that* $r_{l_i} r_{l_j} \sqsubset \mathbf{r}_x$ **then**
        $A^{\mathbf{r}_l}[j, i] = A^{\mathbf{r}_x}[\pi(r_{l_j}, \mathbf{r}_x), \pi(r_{l_i}, \mathbf{r}_x)]$
        continue
      else if $N - l + 1 \geq q$ **then**
        for $z = l$ *to* N **do**
          Calculate $\pi(r_{l_i}, \mathbf{r}_z), \pi(r_{l_j}, \mathbf{r}_z)$ by (12)
          Calculate $f(r_{l_i}, r_{l_j}) +=$
          $$\begin{cases} H(\pi(r_{l_i}, \mathbf{r}_z)), & \text{if } i = j \\ H(\pi(r_{l_j}, \mathbf{r}_z) - \pi(r_{l_i}, \mathbf{r}_z)) H(\pi(r_{l_i}, \mathbf{r}_z)), & \text{otherwise} \end{cases}$$
          if $N - z + f(r_{l_i}, r_{l_j}) < q$ **then**
            break
        end
      end
      if $f(r_{l_i}, r_{l_j}) \geq q$ **then**
        $A^{\mathbf{r}_l}[j, i] = 1$
    end
  end
  Calculate $\kappa_1^{\mathbf{r}_l}(q), \kappa_2^{\mathbf{r}_l}(q)$ by (16) and (17)
  Get $\mathbf{S}_1^{\mathbf{r}_l}(q), \mathbf{S}_2^{\mathbf{r}_l}(q)$ based on $\mathbf{A}^{\mathbf{r}_l}$
end
Calculate $\bar{\kappa}_1(q), \bar{\kappa}_2(q)$ by (8) and (9)
Get $\mathbf{S}_1(q), \mathbf{S}_2(q)$ by $\mathbf{S}_1(q) = \underset{\mathbf{r}_l \in \mathcal{R}}{\cup} \mathbf{S}_1^{\mathbf{r}_l}(q)$, $\mathbf{S}_2(q) = \underset{\mathbf{r}_l \in \mathcal{R}}{\cup} \mathbf{S}_2^{\mathbf{r}_l}(q)$
return $\{\kappa_1^{\mathbf{r}_l}(q), \kappa_2^{\mathbf{r}_l}(q), l = 1, 2, \cdots, N; \mathbf{S}_1^{\mathbf{r}_l}(q), \mathbf{S}_2^{\mathbf{r}_l}(q); \bar{\kappa}_1(q), \bar{\kappa}_2(q); \mathbf{S}_1(q), \mathbf{S}_2(q)\}$

4. Quantifying consensus with consideration of positions and position gaps

The rank positions of an item and the position gaps of pairwise items may be significantly different in a ranking set. Consider the items a and f in Example 1. The rank positions of item a are $\pi(a,\mathbf{r}_1)=1, \pi(a,\mathbf{r}_2)=6, \pi(a,\mathbf{r}_3)=3, \pi(a,\mathbf{r}_4)=2$ and the position gaps of the two items are $\pi(f,\mathbf{r}_1)-\pi(a,\mathbf{r}_1)=5, \pi(f,\mathbf{r}_3)-\pi(a,\mathbf{r}_3)=3, \pi(f,\mathbf{r}_4)-\pi(a,\mathbf{r}_4)=3$. These differences influence the ranking consensus. However, the consensus scores defined in the previous section only involve the existence of q-support patterns. To reflect the importance of these position and gap information, the following definition presents an extension to (6) and (7) for quantifying consensus of a ranking set more effectively.

Definition 4 (Weighted individual consensus scores)**.** *The weighted consensus scores of ranking $\mathbf{r}_l \in \mathcal{R}$ are*

$$\kappa_1^{\mathbf{r}_l}(q) = \frac{1}{N_1^{\mathbf{r}_l}} \sum_{r_{l_i} \in \mathbf{S}_1^{\mathbf{r}_l}(q)} \gamma^{h(r_{l_i},\mathbf{r}_l)} \tag{18}$$

$$\kappa_2^{\mathbf{r}_l}(q) = \frac{1}{N_2^{\mathbf{r}_l}} \sum_{r_{l_i}r_{l_j} \in \mathbf{S}_2^{\mathbf{r}_l}(q)} \lambda^{d(r_{l_i},r_{l_j},\mathbf{r}_l)}, \tag{19}$$

where the constants $0<\gamma\leq 1$ and $0<\lambda\leq 1$ are the weights, $h(r_{l_i},\mathbf{r}_l)$ is the deviation of the position of r_{l_i} in $\mathbf{r}_l$ from its average position in the ranking set, and $d(r_{l_i},r_{l_j},\mathbf{r}_l)$ is the deviation of the position gaps between r_{l_i} and r_{l_j} in $\mathbf{r}_l$ from the average.

The deviations $h(r_{l_i},\mathbf{r}_l)$ and $d(r_{l_i},r_{l_j},\mathbf{r}_l)$ are calculated as follows. For ranking $\mathbf{r}_l \in \mathcal{R}$, we have the sets $\mathbf{S}_1^{\mathbf{r}_l}(q)$ and $\mathbf{S}_2^{\mathbf{r}_l}(q)$ of the q-support patterns defined as (4) and (5), the function $f(r_{l_i},r_{l_j})$ in the form of (14), and the subset $\mathcal{R}'(r_{l_i},r_{l_j})$ of $\mathcal{R}$ containing pattern $r_{l_i}r_{l_j}$ as (1). The average position of item r_{l_i} in the ranking set is defined as

$$\bar{\pi}(r_{l_i}) = \frac{1}{f(r_{l_i},r_{l_i})} \sum_{\mathbf{r}_z \in \mathcal{R}'(r_{l_i},r_{l_i})} \pi(r_{l_i},\mathbf{r}_z). \tag{20}$$

The deviation $h(r_{l_i},\mathbf{r}_l)$ is

$$h(r_{l_i},\mathbf{r}_l) = |\pi(r_{l_i},\mathbf{r}_l) - \bar{\pi}(r_{l_i})|. \tag{21}$$

The position gap between r_{l_i} and r_{l_j} in ranking $\mathbf{r}_z$ is

$$\omega(r_{l_i}, r_{l_j}, \mathbf{r}_z) = \pi(r_{l_j}, \mathbf{r}_z) - \pi(r_{l_i}, \mathbf{r}_z). \tag{22}$$

The average position gap of r_{l_i} and r_{l_j} in the ranking set is defined as

$$\bar{\omega}(r_{l_i}, r_{l_j}) = \frac{1}{f(r_{l_i}, r_{l_j})} \sum_{\mathbf{r}_z \in \mathcal{R}'(r_{l_i}, r_{l_j})} \omega(r_{l_i}, r_{l_j}, \mathbf{r}_z). \tag{23}$$

The deviation $d(r_{l_i}, r_{l_j}, \mathbf{r}_l)$ is

$$d(r_{l_i}, r_{l_j}, \mathbf{r}_l) = |\omega(r_{l_i}, r_{l_j}, \mathbf{r}_l) - \bar{\omega}(r_{l_i}, r_{l_j})|.$$

From the definition, it can be known that smaller values of γ and λ reflect greater impacts of the deviations of item positions and position gaps in rankings on the consensus scores. It is worth noting that the consensus scores defined in the previous section are a special case of the weighted consensus scores with $\gamma = 1, \lambda = 1$. Here, we do not need to make any change to the overall consensus scores defined in Definition 3.

To calculate the weighted consensus scores with the matrix representation, equation (15) in Theorem 1 is changed to

$$A^{\mathbf{r}_l}[j,i] = \begin{cases} \gamma^{h(r_{l_i}, \mathbf{r}_l)}, & \text{if } i = j \text{ and } f(r_{l_i}, r_{l_j}) \geq q \\ \lambda^{d(r_{l_i}, r_{l_j}, \mathbf{r}_l)}, & \text{if } i < j \text{ and } f(r_{l_i}, r_{l_j}) \geq q \\ 0, & \text{otherwise.} \end{cases} \tag{24}$$

Small changes will be needed in Algorithm 1. We follow the steps of Algorithm 1 and change the way to calculate $A^{\mathbf{r}_l}[j,i]$ on Line 8 to the following form

$$A^{\mathbf{r}_l}[j,i] = \begin{cases} H\left(A^{\mathbf{r}_x}[\pi(r_{l_j}, \mathbf{r}_x), \pi(r_{l_i}, \mathbf{r}_x)]\right) \gamma^{h(r_{l_i}, \mathbf{r}_l)}, & \text{if } i = j \\ H\left(A^{\mathbf{r}_x}[\pi(r_{l_j}, \mathbf{r}_x), \pi(r_{l_i}, \mathbf{r}_x)]\right) \lambda^{d(r_{l_i}, r_{l_j}, \mathbf{r}_l)}, & \text{if } i < j. \end{cases}$$

Line 19 is replaced by

$$A^{\mathbf{r}_l}[j,i] = \begin{cases} \gamma^{h(r_{l_i}, \mathbf{r}_l)}, & \text{if } i = j \\ \lambda^{d(r_{l_i}, r_{l_j}, \mathbf{r}_l)}, & \text{if } i < j, \end{cases}$$

and meanwhile the average position $\bar{\pi}(r_{l_i})$ and the average position gap $\bar{\omega}(r_{l_i}, r_{l_j})$ are recorded in here for further use on Line 8.

Remark 1 (Rankings with ties)**.** *Rankings with ties are used in the case that the preferences over some items are identical. Let* $\mathbf{r}_z = (\mathcal{T}_{z_1}, \mathcal{T}_{z_2}, \cdots, \mathcal{T}_{z_n})$ *be a ranking with ties, where* $\mathcal{T}_{z_i}, i \in [1, n]$ *is a set of items with identical preference. For* $i < j$*, every item in* $\mathcal{T}_{z_i}$ *is more preferred than all the items in* $\mathcal{T}_{z_j}$*. The proposed approach can be extended to rankings with ties by making a small change to the position function. Specifically, we can replace (12) with*

$$\pi\left(r_{l_i}, \mathbf{r}_z\right) = \begin{cases} p, & \text{if } r_{l_i} \in \mathcal{T}_{z_p} \\ 0, & \text{otherwise} \end{cases}$$

to make the approach applicable to evaluate consensus of rankings with ties.

5. Detecting outliers

The individual consensus scores $\kappa_1^{\mathbf{r}_l}(q)$ and $\kappa_2^{\mathbf{r}_l}(q)$ directly reflect the (weighted) numbers of q-support patterns that $\mathbf{r}_l$ shares with the other rankings in $\mathcal{R}$. For instance, ranking $\mathbf{r}_3$ in Example 1 shares less 3-support patterns with the others, thus it has much lower consensus scores. This can be used to detect outlier rankings, which have low consensus with most rankings. The following outlier detection method is naturally developed from the consensus quantifying approach.

Consider a ranking set $\mathcal{R}$ with overall consensus scores $\bar{\kappa}_1(q)$ and $\bar{\kappa}_2(q)$ for a given q. Define the relative deviations of the individual consensus scores of ranking $\mathbf{r}_l \in \mathcal{R}$ from the overall consensus scores as

$$v_1^{\mathbf{r}_l}(q) = \frac{\kappa_1^{\mathbf{r}_l}(q) - \bar{\kappa}_1(q)}{\bar{\kappa}_1(q)} \tag{25}$$

$$v_2^{\mathbf{r}_l}(q) = \frac{\kappa_2^{\mathbf{r}_l}(q) - \bar{\kappa}_2(q)}{\bar{\kappa}_2(q)}. \tag{26}$$

Note that $v_1^{\mathbf{r}_l}(q) < 0$ and $v_2^{\mathbf{r}_l}(q) < 0$ imply that the ranking $\mathbf{r}_l$ has lower consensus scores than the overall averages. For given constants $\epsilon_1 > 0$ and $\epsilon_2 > 0$, if $v_1^{\mathbf{r}_l}(q) < -\epsilon_1$ or $v_1^{\mathbf{r}_2}(q) < -\epsilon_2$, we regards $\mathbf{r}_l$ as an outlier of the ranking set. The values of ϵ_1, ϵ_2 depend on the specific need for a system.

This outlier detection method can be used to figure out irrelevant rankings in the ranking set and consequently identify the majority of rankings with higher consensus.

It is of great importance in many scenarios, e.g., design of auto-suggestion queries in search engine. It is worth noting that one potential application of the obtained detection method is to improve rank aggregation. Rank aggregation is the task of aggregating the preferences of different agents to generate a final ranking. The outliers of rankings/agents play a negative role in drawing a consensus ranking. Even though many existing studies have been carried out on rank aggregation [40, 9, 5], there is still room to improve aggregated rankings so that the aggregated result is as close to the ground truth as possible. This will be studied in a separate paper.

6. Experimental studies

This section shows how the proposed approach can be used to evaluate consensus for a set of rankings. The source code is available at `https://github.com/zhiweiuu/secs`.

6.1. Analysis of the Mechanical Turk Dots datasets

The Mechanical Turk Dots datasets [30] include four publicly available datasets obtained for four dots tasks. These datasets each contain rankings obtained by 794 to 800 voters over four candidates. Each candidate corresponds to a certain number of random dots. The voters were asked to rank the candidates from those with the least dots to the most. Each task contains candidates with 200, $200+i$, $200+2i$, and $200+3i$ dots, where $i=3,5,7,9$ respectively for the four tasks. Figure 1 shows the proportions of rankings in each dataset with different Spearman's ρ to the ground truth ranking. The values of different Spearman's ρ are distinguished by colors. It can be seen that the proportions of rankings with high Spearman coefficients 0.8 and 1.0 increase from Dataset 1 to Dataset 4, while that with coefficient 0.4 decreases significantly. The ranking consensus degrees seem increasing from Dataset 1 to Dataset 4. We apply the proposed approach to accurately compare these datasets.

The overall consensus scores without weighting are first considered. Since the datasets have complete rankings, i.e., all the candidates under consideration are ranked in the rankings, the consensus scores of the single items satisfy $\bar{\kappa}_1(q)=4$ for all q and

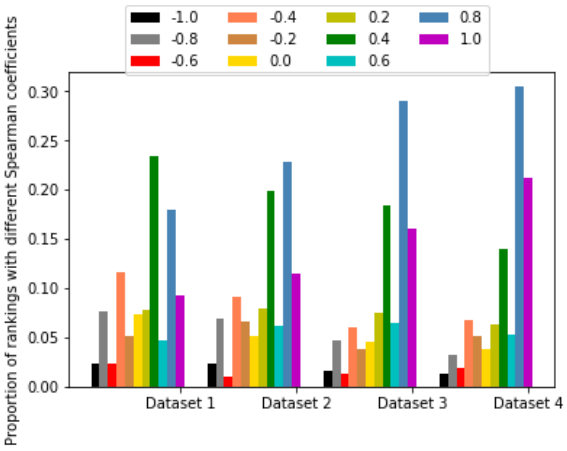


Figure 1: Spearman's ρ between the rankings and the ground truth ranking

all the datasets. Figure 2 gives the overall consensus scores $\bar{\kappa}_2(q)$ with respect to $\frac{q}{N}$, where $\frac{q}{N} \geq 0.5$ indicates that the commonality embedded in half or more than half of the rankings is evaluated. The trend of the overall consensus scores for the four datasets is clear. Dataset 4 has the largest overall consensus score, which indicates that Dataset 4 has the most q-support common patterns. Specifically, it can be seen from the figure that, when $\frac{q}{N}$ is 0.5, the consensus score $\bar{\kappa}_2(q)$ is $0.59, 0.62, 0.68$, and 0.71 respectively for Dataset 1, 2, 3 and 4. This means that on average, $59.00\%, 62.00\%, 68.00\%$, and 71.00% of the pairwise patterns of a ranking are $\lceil \frac{N}{2} \rceil$-support patterns in Dataset 1, 2, 3 and 4, respectively. As the value of q increases, the consensus scores decrease. When $\frac{q}{N}$ reaches 0.67, the consensus score is zero for Dataset 1, which means that arbitrary $q \geq 0.67N$ rankings in the dataset have no common pattern. On the other hand, the consensus scores are 0.12, 0.37, 0.38 for Dataset 2, 3, 4. In other words, on average, $12.00\%, 37.00\%, 38.00\%$ of the patterns of a ranking are supported by at least $\lceil 0.67N \rceil$ rankings in the corresponding dataset.

The overall consensus scores with weightings are then evaluated. Figure 3 shows the consensus scores with respect to the weights γ and λ for a fixed $q = \lceil \frac{N}{2} \rceil$. As shown, $\bar{\kappa}_1(\lceil \frac{N}{2} \rceil)$ and $\bar{\kappa}_2(\lceil \frac{N}{2} \rceil)$ decrease with the increase of weightings on the deviations of positions and position gaps. Dataset 1 has the lowest overall consensus scores and

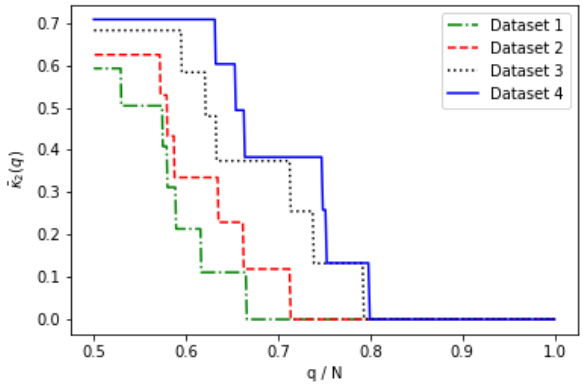


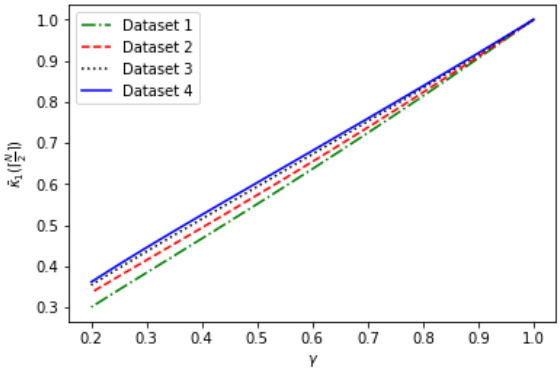
Figure 2: Consensus scores $\bar{\kappa}_2(q)$ of Dots datasets without weighting

Dataset 4 has the highest. The ratios of the consensus scores between Dataset 4 and Dataset 3, Dataset 3 and Dataset 2, and Dataset 2 and Dataset 1 are shown in Table 1 for the cases without weighting and with weighting parameters $\gamma = 0.5, \lambda = 0.5$. By comparing the two cases, it can be found that the ratios with weightings on the deviations of the position and position gaps are higher than those without weightings. This reveals that the differences of the positions of the single q-support items and the position gaps of the q-support patterns decrease from Dataset 1 to Dataset 4.

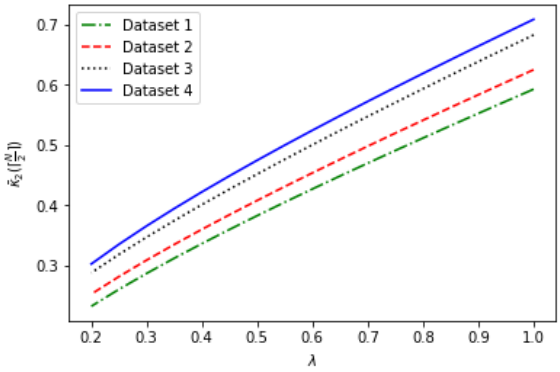
Table 1: Ratios of the consensus scores between datasets

	Dataset 4/Dataset 3	Dataset 3/Dataset 2	Dataset 2/Dataset 1
$\bar{\kappa}_1(\lceil \frac{N}{2} \rceil), \gamma = 1$	1.00	1.00	1.00
$\bar{\kappa}_1(\lceil \frac{N}{2} \rceil), \gamma = 0.5$	1.02	1.04	1.04
$\bar{\kappa}_2(\lceil \frac{N}{2} \rceil), \lambda = 1$	1.04	1.09	1.05
$\bar{\kappa}_2(\lceil \frac{N}{2} \rceil), \lambda = 0.5$	1.05	1.11	1.07

The relative deviations of $\kappa_2^{\mathbf{r}_l}(\lceil \frac{N}{2} \rceil)$ from the overall consensus score $\bar{\kappa}_2(\lceil \frac{N}{2} \rceil)$ is also studied to verify the effectiveness of the proposed outlier detection method. By choosing $\lambda = 0.5$, the result in Table 2 can be obtained. The deviations are very high for $\mathbf{r}_{21}, \mathbf{r}_{24}, \mathbf{r}_{22}, \mathbf{r}_{13}$ of Dataset 1, $\mathbf{r}_{20}, \mathbf{r}_{22}, \mathbf{r}_{19}, \mathbf{r}_{17}$ of Dataset 2, $\mathbf{r}_{19}, \mathbf{r}_{20}, \mathbf{r}_{22}, \mathbf{r}_{15}$ of Dataset 3, and $\mathbf{r}_{21}, \mathbf{r}_{22}, \mathbf{r}_{24}, \mathbf{r}_{20}$ of Dataset 4. These rankings are regarded as out-



(a) $\bar{\kappa}_1(\lceil \frac{N}{2} \rceil)$ with respect to γ



(b) $\bar{\kappa}_2(\lceil \frac{N}{2} \rceil)$ with respect to λ

Figure 3: Weighted consensus scores of Dots datasets

liers of the datasets. They are $(4, 3, 2, 1)$, $(4, 3, 1, 2)$, $(4, 2, 3, 1)$, $(3, 4, 2, 1)$ respectively in each dataset. Note that Spearman's ρ between $(4, 3, 2, 1)$ and the ground truth $(1, 2, 3, 4)$ are -1, and all the Spearman coefficients of the rest three to the ground truth are -0.8. After deleting these outlier rankings, the consensus scores $\bar{\kappa}_1(\lceil \frac{N}{2} \rceil)$ with $\gamma = 0.5$ increase from $0.55, 0.57, 0.59, 0.60$ to $0.58, 0.59, 0.61, 0.62$ for Dataset 1, 2, 3, 4, respectively. The consensus scores $\bar{\kappa}_2(\lceil \frac{N}{2} \rceil)$ change from $0.38, 0.41, 0.45, 0.47$ to $0.42, 0.44, 0.48, 0.49$ for the four datasets. This confirms the effectiveness of the proposed outlier detection method.

It is further found that the four datasets have the same set of the $\lceil \frac{N}{2} \rceil$-support pat-

Table 2: Deviations of the consensus scores

$q = \lceil \frac{N}{2} \rceil$	Dataset 1	Dataset 2	Dataset 3	Dataset 4
$v_2^{\mathbf{r}_1}(q)$	0.72	0.68	0.56	0.55
$v_2^{\mathbf{r}_2}(q)$	0.14	0.38	0.29	0.20
$v_2^{\mathbf{r}_3}(q)$	0.44	0.29	0.28	0.18
$v_2^{\mathbf{r}_4}(q)$	0.44	0.36	0.15	0.08
$v_2^{\mathbf{r}_5}(q)$	0.09	0.08	-0.03	-0.09
$v_2^{\mathbf{r}_6}(q)$	0.06	0.06	-0.03	-0.21
$v_2^{\mathbf{r}_7}(q)$	0.47	-0.01	-0.15	-0.10
$v_2^{\mathbf{r}_8}(q)$	0.11	-0.27	-0.04	-0.11
$v_2^{\mathbf{r}_9}(q)$	0.02	-0.04	-0.35	-0.41
$v_2^{\mathbf{r}_{10}}(q)$	-0.22	0.02	-0.37	-0.21
$v_2^{\mathbf{r}_{11}}(q)$	-0.29	-0.28	-0.17	-0.20
$v_2^{\mathbf{r}_{12}}(q)$	-0.38	-0.11	-0.42	-0.41
$v_2^{\mathbf{r}_{13}}(q)$	**-0.71**	-0.13	-0.20	-0.56
$v_2^{\mathbf{r}_{14}}(q)$	-0.25	-0.15	-0.49	-0.35
$v_2^{\mathbf{r}_{15}}(q)$	-0.06	-0.46	**-0.74**	-0.21
$v_2^{\mathbf{r}_{16}}(q)$	-0.09	-0.45	-0.20	-0.54
$v_2^{\mathbf{r}_{17}}(q)$	-0.36	**-0.72**	-0.28	-0.49
$v_2^{\mathbf{r}_{18}}(q)$	-0.39	-0.37	-0.50	-0.47
$v_2^{\mathbf{r}_{19}}(q)$	-0.40	**-0.74**	**-1.00**	-0.55
$v_2^{\mathbf{r}_{20}}(q)$	-0.46	**-1.00**	**-0.75**	**-0.74**
$v_2^{\mathbf{r}_{21}}(q)$	**-1.00**	-0.44	-0.50	**-1.00**
$v_2^{\mathbf{r}_{22}}(q)$	**-0.73**	**-0.75**	**-0.77**	**-0.75**
$v_2^{\mathbf{r}_{23}}(q)$	-0.12	-0.45	-0.52	-0.54
$v_2^{\mathbf{r}_{24}}(q)$	**-0.74**	-0.47	-0.52	**-0.77**

terns $\mathbf{S}_2(\lceil \frac{N}{2} \rceil) = \{12, 13, 14, 23, 24, 34\}$. By aggregating these $\lceil \frac{N}{2} \rceil$-support patterns, we can obtain the ranking $(1, 2, 3, 4)$, i.e., the ground truth ranking. This enhances the advantage of the proposed consensus quantifying approach over the rank correlation and distance functions approaches, where no common patterns of the rankings are specified.

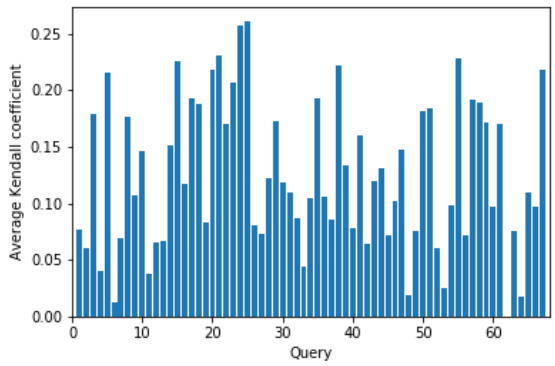
6.2. Evaluation of the information retrieval results of the 2015 CLEFeHealth Lab Task 2

This experiment focuses on top-k rankings using the dataset of the CLEF 2015 eHealth Evaluation Lab Task 2 [32], instead of the complete rankings as in the previous section. The CLEF 2015 eHealth Evaluation Lab Task 2 aimed to foster the design of web search engines in providing access to medical information especially for self-diagnosis information, since commercial search engines were far from being effective in the field. The problem considered in the task was to retrieve web pages for queries related to different medical conditions. The queries were pre-generated by showing images and videos of medical conditions to potential users. There were 67 queries selected to be used in the task for 23 medical conditions, among which 22 conditions had three queries and one condition had one query. The queries were first created in English and then translated into several other languages. The document collection made available to the participates for information retrieval contains approximately one million web pages on a broad range of health topics. The participates were asked to submit up to ten runs for the English queries. The first run of each team was with the highest priority for selection of documents to contribute to the final assessment. Twelve participating teams submitted their English information retrieval results.

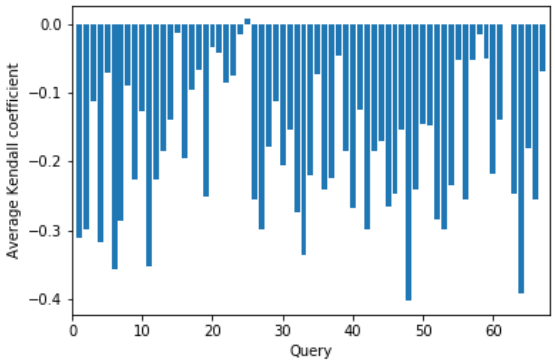
This section evaluates the information retrieval results of the first English runs. Given that the first two pages of a user's search result probably draw the most attention in practice, the top-20 retrieved documents for each query are considered in the evaluation. The conventional Spearman's ρ and Kendall's τ measure the correlation of two complete rankings, as they compare the positions of same items in two rankings. For this dataset with incomplete rankings, the Spearman's ρ and Kendall's τ for top-k rankings proposed in [15] are employed to measure the correlations for the 67

queries. Because a typo exists in the 62^{nd} query, there is no record of some teams for this query in the dataset. This query is not considered in the following analysis. Since there is no ground truth ranking available, we pairwisely compare the ranking obtained by each team for a specific query with the rankings of the other teams and take the average as the comparison result of the team. The obtained comparison results of all the teams for the query are further aggregated by taking their average, and the aggregated result measures the correlations of rankings obtained by all the teams for the query. Figure 4 gives the results of Kendall's τ. Note that a key parameter p is introduced in the calculation of Kendall's τ for top-k rankings in [15]. This parameter corresponds to the penalty for the case that two items σ_i and σ_j appear in one ranking $\mathbf{r}_l$ and none of them are considered in the other compared ranking $\mathbf{r}_z$. In this case, the term $\text{sgn}(\pi(\sigma_i, \mathbf{r}_l) - \pi(\sigma_j, \mathbf{r}_l))\text{sgn}(\pi(\sigma_i, \mathbf{r}_z) - \pi(\sigma_j, \mathbf{r}_z))$ is set to be p. We normalize Kendall's τ to the domain of $[-1, 1]$. The parameter $p = 1$ gives an optimistic approach. It implies that σ_i and σ_j in $\mathbf{r}_z$ are regarded as in the same order as in $\mathbf{r}_l$ when there is no enough information about them. When $p = 0$, it gives a neutral approach. It can be found in Figure 4(a) and Figure 4(b) that Kendall's coefficients are highly depends on the value of p. The result of Spearman's ρ is shown in Figure 5. If an item σ_i in one top-k ranking $\mathbf{r}_l$ does not appear in the other compared top-k ranking $\mathbf{r}_z$, then the position $\pi(\sigma_i, \mathbf{r}_z)$ is set to ℓ. In Figure 5, ℓ is chosen to be $k + 1$. Spearman's ρ also depends on the value of ℓ.

Unlike the Spearman's ρ and Kendall's τ for top-k rankings, where assumptions about unknown factors are made without sufficient information and may consequently lead to bias in the measurement results, the proposed approach has no such problem and the consensus of a ranking set is measured more intuitively based on q-support patterns. It provides a clear understanding about the commonality emmbedded in the rankings obtained with different information retrieval approaches, and it can help to find hard topics in the information retrieval task. Figure 6 shows the 6-support (i.e., $\frac{N}{2}$-support) consensus scores without weightings for the ranking sets of the 66 queries obtained by the 12 teams. The relative values of the consensus scores are generally consistent with the results in Figures 4 and 5. However, our results based on q-support patterns, especially the pairwise patterns, reveal more obvious and detailed information. It can



(a) Optimistic approach



(b) Neutral approach

Figure 4: Average Kendall's τ for the ranking sets of the 66 queries obtained by the 12 teams

be seen from Figure 6(a) that the consensus score $\bar{\kappa}_1(6)$ is greater than 0.5 for queries 10, 13, 15, 20, 24, 25, 31, 38, 57, 58, 59, 67. This means that, on average, more than 50% of the ranked items in a ranking for these queries are emmbedded in at least half of the ranking set. When the orders of these ranked items are further considered, Figure 6(b) shows that, on average, more than 15% of the pairwise patterns of a ranking are supported by at least half of the rankings for queries 20, 24, 25, 38, 57, 58, 59, 67. Figure 7 shows the 6-support consensus scores with the weighting parameters on the deviations of positions and position gaps being $\gamma = 0.9, \lambda = 0.9$. It can be noticed

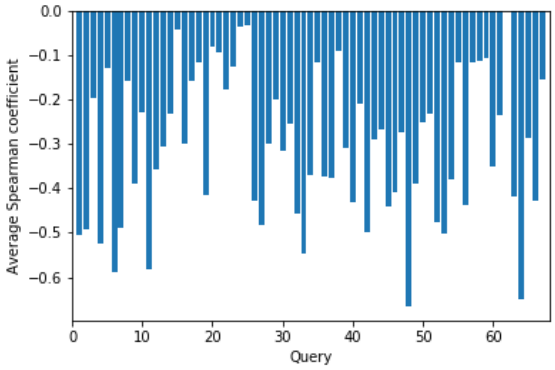


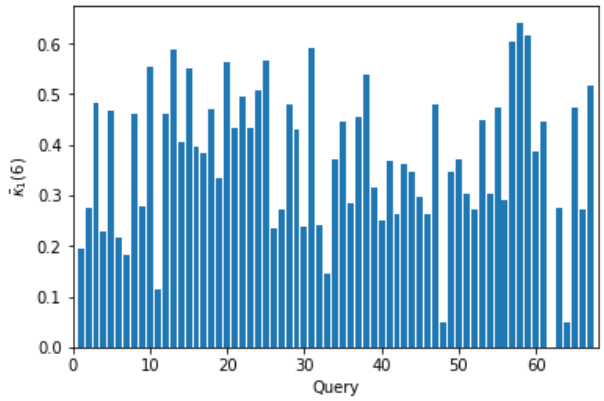
Figure 5: Average Spearman's ρ for the ranking sets of the 66 queries obtained by the 12 teams

that queries 58, 25, 24, 55 have higher consensus scores $\bar{\kappa}_2(6)$, which indicates that the rankings of these queries share more weighted pairwise q-support patterns. Moreover, the consensus scores $\bar{\kappa}_1(6)$ for these queries are also high. In contrast, the consensus scores of queries 64, 48, 11, 33 are much lower. The detailed information of these queries is given in Table 3 and Table 4. By comparing the two tables, it can be found that the queries with clear descriptions or for typical symptoms tend to have higher consensus scores, while vague descriptions or uncommon symptoms lead to retrieval results with lower consensus scores.

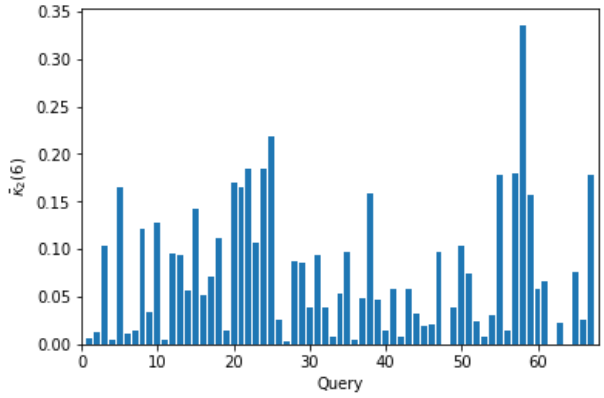
Table 3: Queries with higher consensus scores

Query ID	Query	$\bar{\kappa}_1(6)$	$\bar{\kappa}_2(6)$
58	39 degree and chicken pox	0.47	0.27
25	red rash baby face	0.45	0.17
24	yellow gunk coming from one eye itchy	0.42	0.15
55	crate type mark in skin	0.40	0.15

The consensus of the information retrieval results for each topic is also evaluated with the proposed approach. The queries for each topic are supposed to link to an identical medical condition. The consensus based on 2-support patterns is studied for the 22 topics each with three queries. Topic 13 is not considered, since it associates with



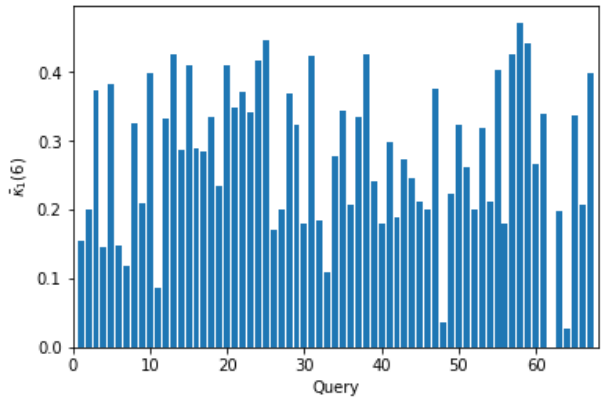
(a) Consensus score $\bar{\kappa}_1(6)$



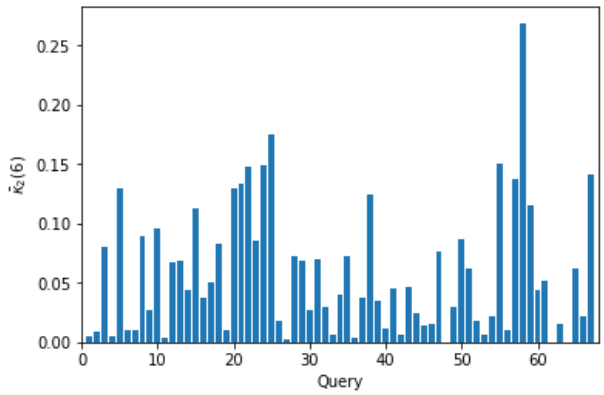
(b) Consensus score $\bar{\kappa}_2(6)$

Figure 6: Consensus scores without weighting for the ranking sets of the 66 queries obtained by the 12 teams

query 62 having incomplete record in the dataset. We take the average of the consensus scores of the ranking sets of the 12 teams. The results are given in Figure 8. Specially, the rankings of topics 15 and 11 have the highest average consensus scores, and the average consensus scores for topic 21 and topic 18 are the lowest. By comparing the topics and the details of the related queries in Table 5 and Table 6, it can be found that the diseases of topics 15 and 11 are more common diseases to be easily self-diagnosed and the generated queries share more commonality. On the contrary, the topics with low consensus scores have more diverse queries, thus they can be regarded as hard



(a) Consensus score $\bar{\kappa}_1(6)$



(b) Consensus score $\bar{\kappa}_2(6)$

Figure 7: Weighted consensus scores for the ranking sets of the 66 queries obtained by the 12 teams

topics, which can be used in further tasks for the development of more advanced search engines.

Table 4: Queries with lower consensus scores

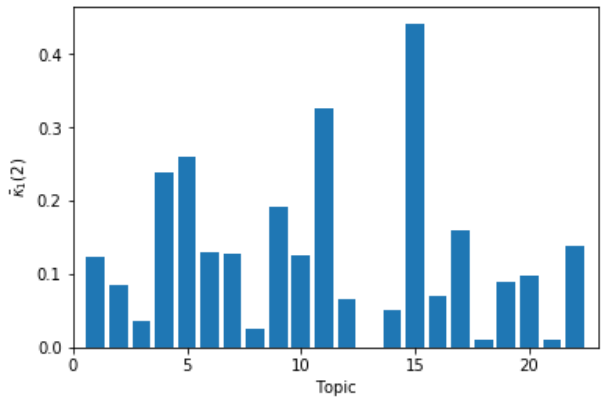
Query ID	Query	$\bar{\kappa}_1(6)$	$\bar{\kappa}_2(6)$
64	involuntary rapid left-right eye motion	0.03	0.00
48	cannot stop moving my eyes medical condition	0.04	0.00
11	white patchiness in mouth	0.09	0.00
33	white infection in pharynx	0.11	0.01

Table 5: Topics with higher average consensus scores

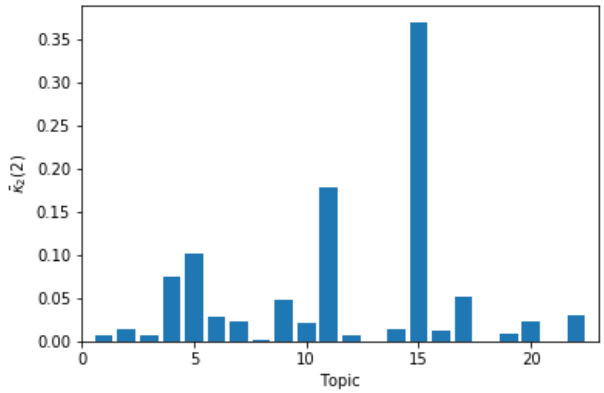
Topic	Query	$\bar{\kappa}_1(2)$	$\bar{\kappa}_2(2)$
15: whooping cough (pertussis)	12: baby has dry cough and has problem to swallow saliva 46: baby cough 66: treatment of coughs in babies	0.44	0.37
11: bronchiolitis (caused by rsv)	31: toddler having squeaky breath 49: baby always breathing with mouth closed 59: heavy and squeaky breath	0.32	0.17

Table 6: Topics with lower average consensus scores

Topic	Query	$\bar{\kappa}_1(2)$	$\bar{\kappa}_2(2)$
21: nystagmus	36: eye are shaking 48: cannot stop moving my eyes medical condition 64: involuntary rapid left-right eye motion	0.01	0.00
18: asthma wheezing	6: child make hissing sound when breathing 15: asthma attack 30: weird sounds when breathing	0.01	0.00



(a) Team average of the consensus score $\bar{\kappa}_1(2)$



(b) Team average of the consensus score $\bar{\kappa}_2(2)$

Figure 8: Average weighted consensus scores for the 22 topics of the 3 queries

7. Conclusion

This paper presents a novel approach to quantifying the consensus degree of a ranking set. A new concept of q-support has been introduced to represent the common patterns embedded in rankings. A matrix representation has been developed to describe the commonality within a ranking set that is shared by an individual ranking, on the basis of which an algorithm has been developed to quantify the consensus efficiently. Moreover, a scheme for detecting outliers in a ranking set is derived from the consensus quantifying approach. Consensus evaluation with weighting on item positions and position gaps has also been considered. Compared with the existing methods based on correlation and distance functions, our approach can characterize and quantify the group preferences more explicitly and it also lays the foundation for the effective detection of outliers and the development of rank aggregation algorithm, which have been illustrated in the experimental studies.

Acknowledgements

This work is supported by the UK EPSRC under Grant No. EP/P031668/1.

References

[1] Alcalde-Unzu J, Vorsatz M. Measuring consensus: Concepts, comparisons, and properties. In: Consensual Processes. Springer; 2011. p. 195–211.

[2] Alcalde-Unzu J, Vorsatz M. Measuring the cohesiveness of preferences: an axiomatic analysis. Social Choice and Welfare 2013;41(4):965–88.

[3] Baigent N. Preference proximity and anonymous social choice. The Quarterly Journal of Economics 1987;102(1):161–9.

[4] Bosch R. Characterizations of voting rules and consensus measures. Ph D Dissertation, Tilburg University 2005;.

[5] Caragiannis I, Chatzigeorgiou X, Krimpas GA, Voudouris AA. Optimizing positional scoring rules for rank aggregation. Artificial Intelligence 2018;.

[6] Carpineto C, Romano G. A survey of automatic query expansion in information retrieval. Acm Computing Surveys (CSUR) 2012;44(1):1.

[7] Carron AV, Brawley LR. Cohesion: Conceptual and measurement issues. Small group research 2000;31(1):89–106.

[8] Carterette B. On rank correlation and the distance between rankings. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM; 2009. p. 436–43.

[9] Chen Y, Suh C. Spectral mle: Top-k rank aggregation from pairwise comparisons. In: International Conference on Machine Learning. 2015. p. 371–80.

[10] Chiniara M, Bentein K. The servant leadership advantage: When perceiving low differentiation in leader-member relationship quality influences team cohesion, team task performance and service ocb. The Leadership Quarterly 2018;29(2):333–45.

[11] Eckert D, Klamler C. Distance-based aggregation theory. In: Consensual processes. Springer; 2011. p. 3–22.

[12] Elzinga C, Wang H, Lin Z, Kumar Y. Concordance and consensus. Information Sciences 2011;181(12):2529–49.

[13] Erdamar B, García-Lapresta JL, Pérez-Román D, Sanver MR. Measuring consensus in a preference-approval context. Information Fusion 2014;17:14–21.

[14] Etesami O, Gohari A. Maximal rank correlation. IEEE Communications Letters 2016;20(1):117–20.

[15] Fagin R, Kumar R, Sivakumar D. Comparing top k lists. SIAM Journal on discrete mathematics 2003;17(1):134–60.

[16] García-Lapresta JL, Pérez-Román D. Consensus measures generated by weighted kemeny distances on weak orders. In: Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on. IEEE; 2010. p. 463–8.

[17] García-Lapresta JL, Pérez-Román D. Measuring consensus in weak orders. In: Consensual processes. Springer; 2011. p. 213–34.

[18] Hashemi V, Endriss U. Measuring diversity of preferences in a group. In: European Conference on Artificial Intelligence. 2014. p. 423–8.

[19] Hassanzadeh FF, Milenkovic O. An axiomatic approach to constructing distances for rank comparison and aggregation. IEEE Transactions on Information Theory 2014;60(10):6417–39.

[20] Henzgen S, Hüllermeier E. Weighted rank correlation: a flexible approach based on fuzzy order relations. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2015. p. 422–37.

[21] Hogg MA. Group cohesiveness: A critical review and some new directions. European review of social psychology 1993;4(1):85–111.

[22] Hotho A, Jäschke R, Schmitz C, Stumme G. Information retrieval in folksonomies: Search and ranking. In: European Semantic Web conference. Springer; 2006. p. 411–26.

[23] Iman RL, Conover W. A measure of top–down correlation. Technometrics 1987;29(3):351–7.

[24] Karpov A. Preference diversity orderings. Group Decision and Negotiation 2017;26(4):753–74.

[25] Kemeny JG. Mathematics without numbers. Daedalus 1959;88(4):577–91.

[26] Kendall MG. A new measure of rank correlation. Biometrika 1938;30(1/2):81–93.

[27] Kim SH, Choi SH, Kim JK. An interactive procedure for multiple attribute group decision making with incomplete information: Range-based approach. European Journal of Operational Research 1999;118(1):139–52.

[28] Kumar R, Vassilvitskii S. Generalized distances between rankings. In: Proceedings of the 19th international conference on World wide web. ACM; 2010. p. 571–80.

[29] Liu TY, et al. Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval 2009;3(3):225–331.

[30] Mao A, Procaccia AD, Chen Y. Better human computation through principled voting. In: Conference on Artificial Intelligence. 2013. .

[31] Nurmi H. A comparison of some distance-based choice rules in ranking environments. Theory and Decision 2004;57(1):5–24.

[32] Palotti JR, Zuccon G, Goeuriot L, Kelly L, Hanbury A, Jones GJ, Lupu M, Pecina P. Clef ehealth evaluation lab 2015, task 2: Retrieving information about medical symptoms. In: CLEF (Working Notes). 2015. p. 1–22.

[33] Poshyvanyk D, Gueheneuc YG, Marcus A, Antoniol G, Rajlich V. Feature location using probabilistic ranking of methods based on execution scenarios and information retrieval. IEEE Transactions on Software Engineering 2007;33(6).

[34] Qin J, Liu X. Multi-attribute group decision making using combined ranking value under interval type-2 fuzzy environment. Information Sciences 2015;297:293–315.

[35] Salas E, Grossman R, Hughes AM, Coultas CW. Measuring team cohesion: Observations from the science. Human factors 2015;57(3):365–74.

[36] Shieh GS. A weighted kendall's tau statistic. Statistics & probability letters 1998;39(1):17–24.

[37] Spearman C. The proof and measurement of association between two things. The American journal of psychology 1904;15(1):72–101.

[38] Tan L, Clarke CL. A family of rank similarity measures based on maximized effectiveness difference. IEEE Transactions on Knowledge and Data Engineering 2015;27(11):2865–77.

[39] Vigna S. A weighted correlation index for rankings with ties. In: Proceedings of the 24th international conference on World Wide Web. International World Wide Web Conferences Steering Committee; 2015. p. 1166–76.

[40] Volkovs MN, Zemel RS. New learning methods for supervised and unsupervised preference aggregation. The Journal of Machine Learning Research 2014;15(1):1135–76.

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

[42] Yilmaz E, Aslam JA, Robertson S. A new rank correlation coefficient for information retrieval. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM; 2008. p. 587–94.

[43] Zhu B, Xu Z, Xu J. Deriving a ranking from hesitant fuzzy preference relations under group decision making. IEEE transactions on cybernetics 2014;44(8):1328–37.