The effect of assessor coverage and assessor accuracy on rank aggregation precision

Laurence A. F. Park and Glenn Stone
Centre for Research in Mathematics
School of Computing, Engineering and Mathematics
University of Western Sydney, Australia
{lapark, g.stone}@uws.edu.au

ABSTRACT

Rank aggregation is the process of aggregating multiple rankings provided by multiple assessors, of a given set of items, into a single ranking. Each assessor, whether it be human or computer based, is a resource that we use to obtain the multiple rankings. The accuracy of the aggregated ranking depends on the accuracy of the assessor ranking and the assessor coverage of the items. Our question is, given limited assessment resources, should each assessor rank many items to obtain item coverage, spending little time on each item, or should each assessor rank only a few items, but spend more time on each item to obtain a high accuracy ranking? In this article, we take a first step towards answering this question, by developing a model, based on simulation, showing the effect of the number of items assigned to an assessor and the accuracy of the assessment on the precision of the aggregated ranking. We find that when using Binomial allocation of items to assessors, increasing the assessor accuracy provides a greater increase in aggregated rank accuracy.

Categories and Subject Descriptors: H.3.3 [Information Search and Retrieval]: Information filtering **General Terms:** Measurement, Experimentation

Keywords: Rank aggregation, assessor accuracy, assessor coverage.

1. INTRODUCTION

Rank aggregation is the process of taking a set of rankings of N items from M assessors and producing a single ranking of the N items. The rankings from the assessors can be complete (each assessors ranks all N items), or incomplete (each assessor ranks a subset of the N items). The rank aggregation process tries to ensure that the rankings of each assessor is reflected in the final ranking, but if there are disagreements amongst the assessors, then the opinions of all assessors cannot be satisfied.

Rank aggregation is commonly used in tournaments, where a set of N items are compared, usually a pair at a time, and an overall ranking must be obtained by the end of the tournament. It is common for sporting tournaments to award points to each pairwise winner, and sum the scores to obtain the overall ranking. The assessors in the Eurovision song competition provide a ranking of their top ten judged songs, where each rank is awarded a predetermined amount of points and the final ranking is obtained by summing the points.

Rank aggregation has been applied where the ranked items are documents, using both manual and automated assessment. Metasearch engines obtain search ranking results from multiple search engines and aggregate the document rankings to obtain a refined ranking. For this case the assessors are the automated search engines. Rank aggregation is also commonly applied to assessment of documents (for conference acceptance, grant awarding or prizes), where submitted documents are manually ranked by a panel and aggregated to obtain an overall ranking.

The purpose of rank aggregation is to reduce the effect of ranking errors from individual assessors. To ensure a high accuracy aggregated ranking, each assessor must provide high accuracy rankings and each item must be assessed by many assessors. It would be ideal for all assessors to assess every item thoroughly. Unfortunately, it is not always feasible to do so, especially when obtaining manual assessments. When assessing a set of academic papers or grant applications, the assessor must invest significant time into understanding the background and contributions of the work in order to provide an accurate ranking. The time contribution also increases with the number of documents that are ranked.

Since assessors generally have limited time for performing the assessment, the question arises "is it better for an assessor to put time and effort in to ranking a small subset of items to obtain accuracy, or should assessors spend the same amount of time assessing a large subset of items to obtain coverage?" (e.g should assessors carefully read three documents, or should they read only the first page of 100 documents?) In other words, how does the number assessors assigned to an item and the assessor accuracy effect the precision of the aggregated ranking or items? In this article, we examine the effect of assessor coverage and assessor accuracy on the accuracy of rank aggregation. The contributions of this article are:

- An analysis of the effect of assessor accuracy and assessor coverage on aggregated rank precision (Section 4),
- A model to describe the relationship between assessor accuracy, assessor coverage and aggregated rank precision (Equation 1),
- Discussion on how the model can be used to examine the effort required by each assessor (Section 5).

The article will proceed as follows: Section 2 provides background on the rank aggregation problem. Section 3 describes the experi-

ADCS, December 08 - 09, 2015, Parramatta, NSW, Australia

ISBN 978-1-4503-4040-3/15/12. . . $15.00

DOI: http://dx.doi.org/10.1145/2838931.2838937

ment performed and data collected. Section 4 provides an analysis of the rank aggregation data and a suggests a fitted model that describes the processes. Finally 5 examines how we can use the model.

2. RANK AGGREGATION

Rank aggregation is the process of combining the information obtained in a set of rankings of a given set of items, to obtain a single ranking of the set of items. Before we can perform rank aggregation, we must obtain the set of rankings from a set of assessors, where each assessor provides either a manual or automatic ranking of an assigned subset of the items. The assessor's task is to examine the assigned subset of items and order them according to a set of criteria, where each assessor is given the same set of criteria. The rank aggregation method is then applied to the set of rankings obtained from each assessor, to obtain an overall ranking of the complete set of items, and hopefully remove ranking errors from assessors.

Using many assessors and aggregating their rankings rather than obtaining an individual ranking: 1) allows us to obtain a distributed and hopefully independent set of rankings, which when combined, assists in reducing the effect of any ranking errors made by individual assessors, and 2) also reduces the work load of each assessor if each assessor ranks a subset of items.

There have been many studies on methods of rank aggregation [4, 2, 1, 6, 7, 8, 9, 3, 10] but little in terms of examining the parameters of the assessors on the aggregated ranking. Rank aggregation methods usually employ the use of high accuracy assessors in order to obtain a high accuracy aggregated rank. But ensemble methods of machine learning such as bagging, boosting and stacking have shown that we can successfully combine the decisions of many low accuracy systems (weak learners) with high coverage across systems to obtain accurate predictions [5].

We will examine how the assessor parameters of Assessor accuracy and Assessor coverage effect the precision of the aggregated ranking. This relationship can be used to determine how we can best assign items to assessors and how to instruct them assessors on how thorough their ranking should be. To examine the relationship between precision and work, we define the following variables:

- **Precision at** k: The proportion of items ranked in the top k of the aggregated ranking that are also in the top k of the true ranking.
- **Assessor coverage**: The proportion of assessors assigned to each item.
- **Assessor accuracy**: A measure of the similarity between the set of assessor's rankings and the true rankings of the set.

The first two of these variables are clearly defined as proportions; for the third we will use the expected Kendall's τ rank correlation.

3. EXPERIMENT

To examine the effect of assessor coverage and assessor accuracy on the precision of the aggregated ranking, we will run a simulation and attempt to find a model that explains the relationships between the variables in the simulation. The variables in the experiment are:

- the method of aggregation
- the assessor coverage ($\alpha \in [0, 1]$)
- the assessor accuracy ($\mathbb{E}[\tau] \in [-1, 1]$)
- the number of items being assessed ($N \in \mathbb{N}^+$)
- the number of assessors ($M \in \mathbb{N}^+$)

We kept constant the number of items being assessed ($N = 1000$) and the number of assessors ($M = 500$). These two values were chosen to be of similar value to the numbers found in grant awarding systems and large conference review systems.

To simulate the ranking process:

1. Generate a list of N items with true rank $\mathcal{R}$ from 1 to N.
2. The N items are assigned to M groups, where each group represents the subset to be ranked by an assessor and each item is assigned to αM groups.
3. For each group of items assigned to an assessor, we induce a ranking of the items with expected Kendall's tau correlation $\mathbb{E}[\tau]$ to the true ranking of the subset of items. If $\mathbb{E}[\tau] = 1$, then there is no error in the assessor's ranking. If $\mathbb{E}[\tau] < 1$, then there may be error.
4. The set of M partial ranked lists are then aggregated using the chosen aggregation algorithm to form the complete ranking R of the N items.
5. The precision at k is computed as the proportion of top k ranked items from R that are also in the top k ranked items of $\mathcal{R}$.

This process was run using $\mathbb{E}[\tau] = 0.1$ to 1 in increments of 0.1 and $\alpha = 0.01, 0.02, 0.04, 0.1, 0.2, 0.4$ and 1, corresponding to 5, 10, 20, 50, 100, 200 and 500 assessors per item. The rank aggregation was computed using the methods Borda, SFO, WAPR, MC1, MC2, MC3 and MC4. All of these methods except WAPR are described in [4]. Weighted average precision rank (WAPR) was the previous rank aggregation method used by the Australian Research Council for ranking national grant applications. WAPR converts the assessor's ranks to percentiles then aggregates rankings as a weighted average of the percentiles. The weight assigned to an assessor's set of percentiles is equal to the number of items ranked, capped at 25.

3.1 Simulating assessor accuracy

To simulate no assessor ranking error, we simply order each item allocated to an assessor in the correct order. A simple method of simulating error in an assessor's ranking is to randomly permute the correct ordering of a given number of adjacent items. Recently [11] investigated methods of inducing error in rankings and found that although random adjacent transpositions is the common form of inducing error, we are unable to easily compute the expected Kendall's τ rank correlation and so we are unable to parameterise and control the error. A new form of rank error induction was developed called Gaussian Perturbation that allows us to randomly permute the ranking of a set of items, given the number of items and the expected Kendall's τ rank correlation $\mathbb{E}[\tau]$. We used Gaussian Perturbation to simulate assessor error and control the level of error using the simulation variable $\mathbb{E}[\tau]$, the exptected Kendall's τ rank correlation.

3.2 Allocation of items to assessors

Our experiment required that each item was assessed by αM assessors, ensuring that each assessor did not receive duplicate items. To assign each item to assessors, we took a random sample of αM assessors from the pool of M assessors, without replacement. The

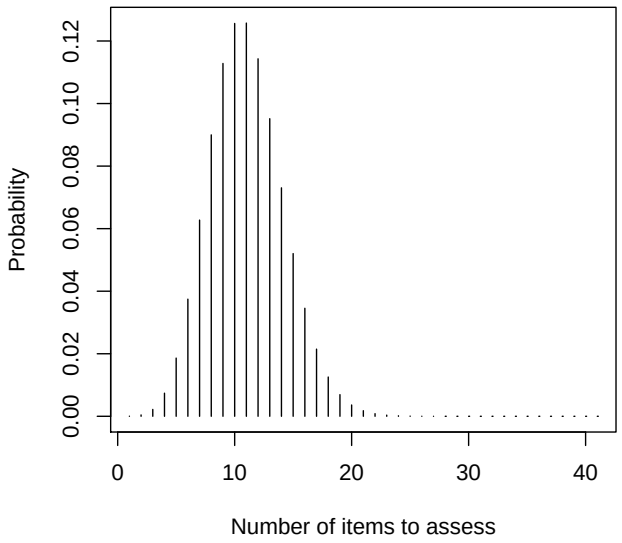

Figure 1: Distribution of the number of items assigned to each assessor for $N = 1000$ items and the probability of each item being assigned to a given assessor is $\alpha = 0.01$.

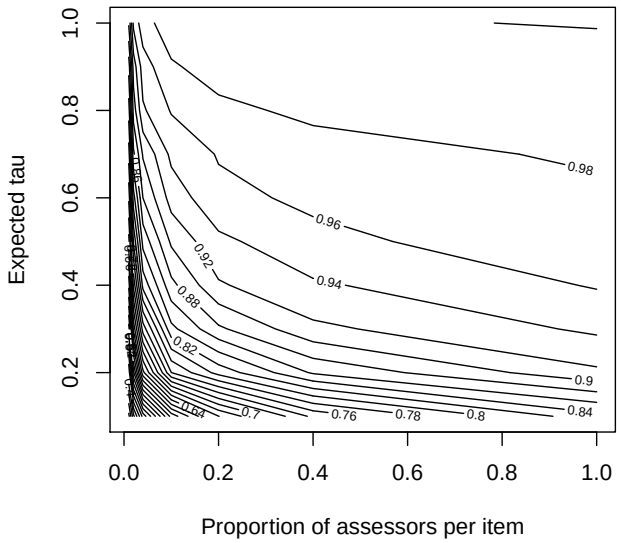

Figure 2: Contour plot of the mean rank aggregation precision.

ith item was then added to the subset of items to assess for each of the αM assessors. This process was performed for all N items.

The probability of a given assessor being chosen to assess item i is α. Therefore the number of items assigned to assessor j follows a Binomial distribution with $n = N$ trials and probability of success $p = \alpha$. The expected number of items to assess per assessor is αN with a standard deviation of $\sqrt{\alpha(1-\alpha)N}$. A discrete density plot of the distribution for $N = 1000$, and $\alpha = 0.01$ is given in Figure 1.

The ten values of $\mathbb{E}[\tau]$, seven values of α and seven aggregation methods gives 490 variable combinations. We also used 10 replicates of the 490 combinations providing us with 4900 measurements of rank aggregation precision. The replicates were required to take into account variance from the random item allocation to assessors and the random rank accuracy simulated for each assessor.

4. ANALYSIS

A contour plot of the simulation results is given in Figure 2 showing the precision at 200 for α and $\mathbb{E}[\tau]$ between 0 and 1. The contour flattens out horizontally and vertically at the lower right and upper left corners of the plot. This implies that as assessor accuracy decreases, the assessor coverage has little impact, and vice versa. To examine how assessor accuracy ($\mathbb{E}[\tau]$) and assessor coverage (α) effect the aggregated rank precision at 200 (p), we will determine a model form, then fit the model and examine the fitted parameters and goodness of fit. The model must adhere to the three properties:

1. If the assessor expected accuracy parameter $\mathbb{E}[\tau] = 0$, then we expect a randomly permuted ranking from each assessor. A random ranking contains no information, so any rank aggregation method would result in an overall random ranking. This implies that the top n items that appear in the aggregated list would be equally as precise as choosing n items out of a hat. Therefore, if $\mathbb{E}[\tau] = 0$, the expected precision at 200 is 0.2, independent of the number of items assigned to each assessor.

2. The assessor coverage parameter α is restricted to the domain $[0, 1]$. Letting $\alpha = 0$ leads to all items being ranked by 0 assessors, meaning we obtain no information from assessors. This implies that we obtain randomly sorted rankings, leading to a random aggregated ranking ($p = 0.2$) independent of the assessor accuracy.

3. If $\alpha = 1$ and $\mathbb{E}[\tau] = 1$, then each assessor supplies a perfect ranking of all N items, so the aggregation method will provide the correct ranking of all items, providing $p = 1$.

After analysis of our data, and using the constraints discussed above, we arrived at the model:

$$\text{logit}\left((p - 0.2)/0.8\right) = \beta_0 + \beta_1 \log(\alpha) + \beta_2 \log(\mathbb{E}[\tau]) \quad (1)$$

or in terms of p:

$$p = \frac{0.8}{1 + e^{-(\beta_0 + \beta_1 \log(\alpha) + \beta_2 \log(\mathbb{E}[\tau]))}} + 0.2 \quad (2)$$

we can see that if β_1 and β_2 are positive, then p increases as $\mathbb{E}[\tau]$ increases or α increases. If $\mathbb{E}[\tau] = 0$ or $\alpha = 0$ then $p = 0.2$, and p is dependent on $\mathbb{E}[\tau]$ and α for all $\mathbb{E}[\tau] > 0$ and $\alpha > 0$. If $\alpha = 1$ and $\mathbb{E}[\tau] = 1$, then $p = 0.8/(1 + e^{-\beta_0}) + 0.2$, therefore if β_0 is large, $p = 1$. Fitting the model to our 4900 simulation points gives the coefficients:

- $\beta_0 = 4.791$ (standard error of 0.0131)
- $\beta_1 = 0.668$ (standard error of 0.00395)
- $\beta_2 = 1.623$ (standard error of 0.00871)

with $R^2 = 0.9281$ showing that $\mathbb{E}[\tau]$ and α explain most of the variance in p, with the remainder of the variance due to the selec-

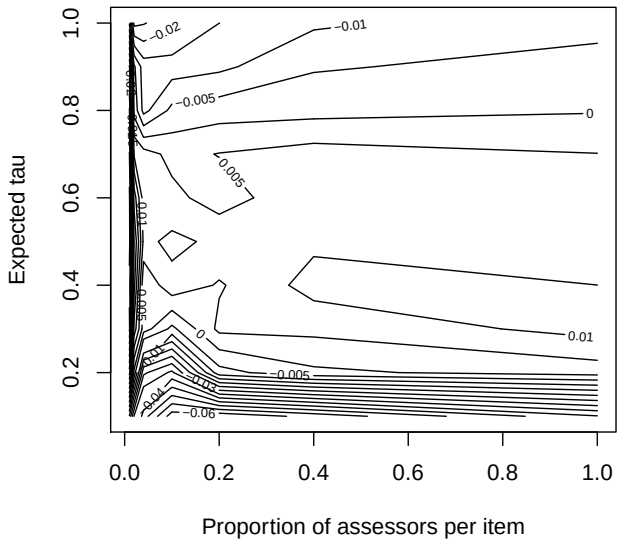

Figure 3: Contour plot of the difference in predicted precision and mean precision.

tion of aggregation method[1] and the random replicates. A variable for the aggregation method was not included in the model because we want to obtain a relation between p, α and $\mathbb{E}[\tau]$, independent of the aggregation method. We also find that the fitted value of β_0 leads to $p = 0.993$ when $\alpha = 1$ and $\mathbb{E}[\tau] = 1$, which is close to the desired value of 1.

Figure 3 shows a contour plot of the model error (predicted precision using the model in equation 2 minus the mean precision from the simulation). We can see the error is small for all values of α and $\mathbb{E}[\tau]$. The greatest error being -0.06 when $\mathbb{E}[\tau]$ is low and α is about 0.2, showing that our proposed model is a good model for explaining effect of α and $\mathbb{E}[\tau]$ on p.

5. DISCUSSION

We began the article asking how does the assessor coverage and the assessor accuracy effect the precision of the aggregated ranking, and we have found a suitable model in equation 2, given Binomial allocation of items to assessors. The model shows that precision increases at a faster rate when increasing $\mathbb{E}[\tau]$ compared to α. By examining the ratios of the fitted coefficients, we find that to obtain the increase in precision from doubling the assessor coverage, we must increase $\mathbb{E}[\tau]$ by only a factor of $2^{\beta_1/\beta_2} = 1.330$.

We asked this question because we were interested in which of the two parameters is best to increase to obtain the greatest increase in precision, regardless of the aggregation method used. To fully answer this question, we must also take into account the assessor effort required to increase these parameters. We believe it is valid to assume that assessor effort is linearly proportional to the α, the number of items assessed. However, it is not clear how effort is related to $\mathbb{E}[\tau]$. $\mathbb{E}[\tau]$ is a measure of accuracy, therefore it should require greater effort to increase at high values, therefore effort may be exponentially related to $\mathbb{E}[\tau]$, with $\mathbb{E}[\tau] = 0$ requiring zero effort and an asymptote at $\mathbb{E}[\tau] = 1$. To investigate this relationship further, user studies or system studies are required.

[1]Note that including the aggregation method as a factor increases R^2 to 0.9332.

6. CONCLUSION

When a ranking of items is required, but it is not feasible for one assessor to rank the set (due to time constraints or accuracy), we resort to employing the use of many assessors, where each assessor is given a subset of the items to rank and the ranked subsets are aggregated to obtain the complete ranking.

Rank aggregation methods depend on assessor accuracy and assessor coverage. It is common for assessors to be given a small subset of the items to rank, where they are asked to rank the items to the best of their ability in a given time, but this approach has little assessor coverage.

In this article, we investigated the relationship between aggregated rank precision, assessor accuracy and assessor coverage. To identify this relationship, we simulated the aggregation process using Binomial allocation over a set of parameters and found a suitable model for the relationship between the number of items assigned to an assessor, the assessor accuracy and the aggregated rank precision.

We came to the conclusion that, when 1000 items are Binomially allocated to 500 assessors, we achieve a greater increase in precision by increasing the assessor accuracy, but further analysis is needed to determine the effort required to achieve an increase in assessor accuracy.

References

[1] N. Ailon. Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica*, 57(2):284–300, 2010.

[2] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):23, 2008.

[3] R. Arora and M. Meila. Consensus ranking with signed permutations. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 117–125, 2013.

[4] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.

[5] J. Friedman, T. Hastie, R. Tibshirani, et al. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2):337–407, 2000.

[6] D. F. Gleich and L.-h. Lim. Rank aggregation via nuclear norm minimization. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 60–68. ACM, 2011.

[7] J. Guiver and E. Snelson. Bayesian inference for plackett-luce ranking models. In *proceedings of the 26th annual international conference on machine learning*, pages 377–384. ACM, 2009.

[8] D. S. Hochbaum. Ranking sports teams and the inverse equal paths problem. In *Internet and Network Economics*, pages 307–318. Springer, 2006.

[9] A. Klementiev, D. Roth, and K. Small. Unsupervised rank aggregation with distance-based models. In *Proceedings of the 25th international conference on Machine learning*, pages 472–479. ACM, 2008.

[10] S. Negahban, S. Oh, and D. Shah. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2012.

[11] L. A. F. Park and G. Stone. Inducing controlled error over variable length ranked lists. In *Advances in Knowledge Discovery and Data Mining*, volume 8444 of *Lecture Notes in Computer Science*, pages 259–270. Springer International Publishing, 2014.