A Head-Weighted Gap-Sensitive Correlation Coefficient

Ning Gao
College of Information Studies/UMIACS
University of Maryland, College Park
ninggao@umd.edu

Douglas W. Oard
College of Information Studies/UMIACS
University of Maryland, College Park
oard@umd.edu

ABSTRACT

Information retrieval systems rank documents, and shared-task evaluations yield results that can be used to rank information retrieval systems. Comparing rankings in ways that can yield useful insights is thus an important capability. When making such comparisons, it is often useful to give greater weight to comparisons near the head of a ranked list than to what happens further down. This is the focus of the widely used τ_{AP} measure. When scores are available, gap-sensitive measures give greater weight to larger differences than to smaller ones. This is the focus of the widely used Pearson correlation measure (ρ). This paper introduces a new measure, τ_{GAP}, which combines both features. System comparisons from the TREC 5 Ad Hoc track are used to illustrate the differences in emphasis achieved by τ_{AP}, ρ, and the proposed τ_{GAP}.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Performance evaluation

Keywords

Evaluation Metric; Rank Correlation Coefficient

1. INTRODUCTION

In information retrieval evaluation, we often wish to compare alternative systems based on some single-valued evaluation measures. For example, we might want to know whether comparing systems using relevance judgments created by one user (to whom we have access) can be used to compare systems in ways that are predictive of what we would have seen had some other user made the judgments [9]. Alternatively, we might want to know whether we can better approximate the system comparison results we could compute with very extensive relevance judgments on a large number of topics by reducing the number of topics or by reducing the number of judgments per topic [2]. In such cases, we can formulate our research question as asking about the correlation between two ranked lists of scores, where the scores result from some evaluation metric such as F_1, Expected Reciprocal Rank (ERR), Normalized Discounted Cumulative Gain (NDCG), or Mean Average Precision (MAP).

When making such comparisons, we focus on two desiderata. First, we prefer that those differences to be large, since small differences may not reflect any meaningful degree of impact on the user experience [3]. Second, we prefer that those differences be statistically significant, since unreliable measurements of differences could be misleading [7]. When extending our comparison from pairs to large sets of systems, as is common in shared-task evaluations such as TREC, CLEF, NTCIR and FIRE, we often care more about distinguishing systems that are very different from each other, which requires the evaluation metric to be gap-sensitive. We also care more about the comparisons among the best systems than we do about comparisons between, for example, the best and the worst systems, which requires the evaluation metric to be head-weighted.

Perhaps the most widely used measure of rank correlation in information retrieval research is Kendall's τ [4] in which swapping systems is penalized. Yilmaz et al. [10] introduced a head-weighted variant of τ that they call τ_{AP} that penalizes the misranking of the best systems (i.e., those near the head of the list), and that measure is now also often reported. Buckley and Voorhees have observed, however, that when systems receive very similar scores we should care less about swaps than when those scores are very different [1]. They therefore created a gap-sensitive measure by suppressing the effect of small swaps by treating any swap within a "fuzziness value" (e.g., 5% relative to the smaller value) as not being large enough to be counted. In this paper, we propose to generalize that approach to penalize swaps in proportion to the difference in scores, so that large swaps will gave the greatest influence on the measure, but small swaps are not completely ignored. This approach thus potentially offers greater insight, without the need to commit to a specific "fuzziness" threshold. Moreover, we combine this more nuanced approach with the head-weighted design of τ_{AP} to produce a new correlation measure for ranked lists of scores that is both head-weighted and gap-sensitive. We call our new measure τ_{GAP} (for Gap And Position).

The remainder of the paper is organized as follows, Section 2 reviews the prior work on the use of correlation measures for system comparison. Section 3 then define τ_{GAP} and establishes that it has a number of desirable properties. Section 4 complements this analytic perspective with empirical results for system scores from the TREC 5 Ad Hoc task, and then further analyses the focus of different correlation coefficient metrics through the heatmap of system pair weights. Section 5 concludes the paper with some remarks on limitations of the τ_{GAP} measure that may inspire further work on this important problem.

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

DOI: http://dx.doi.org/10.1145/2766462.2767793.

2. RELATED WORK

The Pearson correlation coefficient between two items is defined as the covariance of the two items divided by the product of their standard deviations:

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y} \tag{1}$$

where X and Y are the vectors of ranked lists; E is the expectation; σ is the standard deviation; and μ is the mean [6]. Given two ranked lists of items, the Spearman correlation coefficient [11] is defined as the Pearson correlation coefficient between the ranks (i.e., with the ranks used in place of scores). The most widely used measure in information retrieval research is neither Pearson nor Spearman correlation, however, but rather Kendall's τ [4]. Kendall's τ evaluates the correlation of two lists of items by counting their concordant and discordant pairs.

To fulfill the specific evaluation needs for different tasks, various definitions of correlation coefficients derived from Kendall's τ have been proposed. Kendall's tau rank distance [5] measures the disagreements between two ranking lists by counting the swaps that the bubble sort algorithm needed to sort one list in the same order as the other. The *AP Correlation* coefficient (τ_{AP}), proposed by Yilmaz et al. [10], focuses and penalizes more on the errors at high rankings. Given a ground-truth list and prediction list, τ_{AP} of the two lists is defined as:

$$\tau_{AP} = \frac{2}{(N-1)} \cdot \sum_{i=2}^{N} \frac{C_i}{i-1} - 1 \tag{2}$$

where N is the number of items in the list; C_i is the number of items above rank i in the prediction list and correctly ranked with respect to the item at rank i in the ground-truth list. For each item at rank i, τ_{AP} only checks the positions of the $i-1$ items with ranks above i, and calculates the proportion of the correctly ordered items with respect to the item at rank i. Finally, the value of τ_{AP} is a linear combination of all ranks i in the prediction list.

3. THE TAU GAP COEFFICIENT

τ_{GAP} is a non-parametric correlation coefficient, particularly sensitive to errors at high rankings and errors for item pairs with large score differences (gaps). In this section, we present the definition of τ_{GAP}. Since the definition of τ_{AP} and τ_{GAP} are both based on swapped item pairs counting, we further explore the mathematical properties of τ_{GAP} comparing with τ_{AP}.

$$\tau_{GAP} = \frac{2}{(N-1)} \cdot \sum_{i=2}^{N} \frac{\sum_{j<i} |CG_{ji}|}{\sum_{j<i} |G_{ji}|} - 1 \tag{3}$$

Given two ranked lists, a ground truth list and a prediction list, τ_{GAP} is defined as in Equation 3, where N is the number of items in the lists; i represents the item at rank i in the prediction list; j is the item ranked higher than i in the prediction list; G_{ji} is the gap between the item at rank j and item at rank i in ground-truth list; CG_{ji} returns the same value as G_{ji} when the item pair (j,i) in the prediction list are ranked in the same order as in the ground-truth list, otherwise, return 0. For each item at rank i in the prediction list, τ_{GAP} only considers the rank relation of the $i-1$ item pairs (j,i). $\sum_{j<i} |G_{ji}|$ is the sum of all the gaps for the $i-1$ item pairs, and $\sum_{j<i} |CG_{ji}|$ is the sum of the gaps for correctly ordered pairs.

THEOREM 1. *The value of τ_{GAP} is always between -1 and 1.*

PROOF. For each item at rank i, there will be $i-1$ pairs of items (j,i) taken into the consideration by τ_{GAP}. $\sum_{j<i} |G_{ji}|$ is the sum of the gaps of these $i-1$ pairs of items. $\sum_{j<i} |CG_{ji}|$ is the sum of the gaps that the prediction list ranks the two items (j,i) in the same order as the ground-truth list. Therefore, $\frac{\sum_{j<i} |CG_{ji}|}{\sum_{j<i} |G_{ji}|}$ is always between 0 and 1. Normalized across all ranks i, $\frac{1}{(N-1)} \cdot \sum_{i=2}^{N} \frac{\sum_{j<i} |CG_{ji}|}{\sum_{j<i} |G_{ji}|}$ is always between 0 and 1; $\frac{2}{(N-1)} \cdot \sum_{i=2}^{N} \frac{\sum_{j<i} |CG_{ji}|}{\sum_{j<i} |G_{ji}|}$ is always between 0 and 2. Then the value of τ_{GAP} is always between -1 and 1. □

THEOREM 2. *If the gaps between the items follow the uniform distribution, then τ_{GAP} is equal to τ_{AP}.*

PROOF. Let the uniform gap be g, then $\sum_{j<i} |G_{ji}| = (i-1) \cdot g$, and $\sum_{j<i} |CG_{ji}| = C_i \cdot g$. Therefore,

$$\begin{aligned} \tau_{GAP} &= \frac{2}{(N-1)} \cdot \sum_{i=2}^{N} \frac{C_i \cdot g}{(i-1) \cdot g} - 1 \\ &= \frac{2}{(N-1)} \cdot \sum_{i=2}^{N} \frac{C_i}{(i-1)} - 1 = \tau_{AP} \end{aligned} \tag{4}$$

Moreover, if the errors are uniformly distributed over the all the ranks in prediction list, then τ_{GAP}, τ_{AP} and τ are equivalent. □

THEOREM 3. *For two prediction lists with same number of items N and same number of errors located at the same ranks, if the errors of one list have large gaps and the errors of the other list have small gaps, then the value of τ_{GAP} for the list with large error gaps $\tau_{GAP-Large}$ will be smaller than the τ_{GAP} for the list with small error gaps $\tau_{GAP-Small}$.*

PROOF. If we define FG_{ji} as the gap between items at rank i and rank j when the item pair (j,i) in the prediction list are ranked in the converse order as in the ground-truth list, or otherwise 0, then τ_{GAP} could be represented as:

$$\tau_{GAP} = \frac{2}{(N-1)} \cdot \sum_{i=2}^{N} \frac{\sum_{j<i} |CG_{ji}|}{\sum_{j<i} |CG_{ji}| + \sum_{j<i} |FG_{ji}|} - 1 \tag{5}$$

If two prediction lists have the same number of errors located at the same ranks, then their will have the same N, i, j and CG_{ji}. The only difference between $\tau_{GAP-Large}$ and $\tau_{GAP-Small}$ is that $\tau_{GAP-Large}$ has larger values for FG_{ji}, so that the value of $\tau_{GAP-Large}$ will be always smaller than $\tau_{GAP-Small}$. □

THEOREM 4. *For a prediction list, if the swapped item pairs always have small gaps, then the value of τ_{GAP} is larger than τ_{AP}; if the swapped item pairs always have large gaps, then the value of τ_{GAP} is smaller than τ_{AP}.*

PROOF. The expectation of the difference between τ_{GAP} and τ_{AP} is:

$$\begin{aligned} E[\tau_{GAP} - \tau_{AP}] &= \frac{2}{N-1} \cdot \sum_{i=2}^{N} \left(\frac{\sum_{j<i} CG_{ji}}{\sum_{j<i} G_{ji}} - \frac{C_i}{i-1} \right) \\ &= \frac{2}{N-1} \cdot \sum_{i=2}^{N} \frac{(i-1)\sum_{j<i} CG_{ji} - C_i \sum_{j<i} G_{ji}}{(i-1)\sum_{j<i} G_{ji}} \end{aligned} \tag{6}$$

Let P_i be the probability that the items before item i in the prediction list are ranked as the same order as in ground-truth list, then we have

$$C_i = P_i \cdot (i-1) \tag{7}$$

Then

$$E[\tau_{GAP} - \tau_{AP}] = \frac{2}{N-1} \cdot \sum_{i=2}^{N} \frac{(i-1)(\sum_{j<i} CG_{ji} - P_i \sum_{j<i} G_{ji})}{(i-1)\sum_{j<i} G_{ji}}$$
$$= \frac{2}{N-1} \cdot \sum_{i=2}^{N} \left(\frac{1}{\sum_{j<i} G_{ji}} \cdot \left(\sum_{j<i} CG_{ji} - P_i \sum_{j<i} G_{ji} \right) \right) \quad (8)$$

The overall expectation of $E[\tau_{GAP} - \tau_{AP}]$ is a linear combination over all ranks $i \in [2, N]$. However, for each rank position i in the predict list, the expectation of the difference between τ_{GAP} and τ_{AP} is:

$$E_i[\tau_{GAP} - \tau_{AP}] = \frac{2}{\sum_{j<i} G_{ji}} \cdot \left(\sum_{j<i} CG_{ji} - P_i \sum_{j<i} G_{ji} \right) \quad (9)$$

Since $\sum_{j<i} G_{ji}$ is always positive, the relation between τ_{GAP} and τ_{AP} at rank i depends on $(\sum_{j<i} CG_{ji} - P_i \sum_{j<i} G_{ji})$. For rank i, there are $(i-1)$ pairs of items and their corresponding gaps in the prediction list taken into consideration, and there are $P_i(i-1)$ pairs ordered correctly. $\sum_{j<i} CG_{ji}$ is the sum of the gaps for these correct ordered pairs, and $P_i \sum_{j<i} G_{ji}$ is a definite proportion of the total gaps of the $(i-1)$ pairs. If the $P_i(i-1)$ correctly ordered pairs with large gaps, then $\sum_{j<i} CG_{ji} > P_i \sum_{j<i} G_{ji}$, and $E_i[\tau_{GAP} - \tau_{AP}] > 0$; if the $P_i(i-1)$ correctly ordered pairs focus on the pairs with small gaps, then $E_i[\tau_{GAP} - \tau_{AP}] < 0$; ideally, if the $P_i(i-1)$ correctly ordered pairs are distributed randomly across all levels of gaps, then $E_i[\tau_{GAP} - \tau_{AP}] = 0$. Since the expectation of $E[\tau_{GAP} - \tau_{AP}]$ is a linear combination over all $E_i[\tau_{GAP} - \tau_{AP}]$, we could conclude that if the error gaps of a prediction list are relatively small, then τ_{GAP} is larger than τ_{AP}; if the error gaps of a prediction list are relatively large, then τ_{GAP} is smaller than τ_{AP}. □

4. COMPARISON OF THE METRICS

In this section, we use 61 participating systems from TREC 5, ranked by MAP, as a case study to compare the different emphases of the correlation coefficients ρ, τ_{AP} and our proposed τ_{GAP}.

4.1 Correlation Scores for Prediction Lists

Figure 1 shows the result of a two-sided paired t-test for each pair of systems based on Average Precision (AP) scores for each of the 50 topics as samples, with both axes sorted in the same order. In that figure, system pairs with $p < 0.05$ are plotted as white (37%) and system pairs with $p \geq 0.05$ are plotted as black (63%). This illustrates clearly that swaps among many of the systems (near the main diagonal, where score differences are smallest) would offer little insight into whether one evaluation framework yielded system comparisons that we meaningfully different from another.

Figures 2 and 3 then each illustrate two ways of making ranked lists to compare. Each dot represents a system, with the true (TREC-5) rank of that system plotted on the X-axis, and a randomly permuted rank on the Y-axis. We produce the random permutations by randomly selecting five system pairs each time and swapping them. In Figure 2, we randomly select systems pairs that are statistically significantly different from each other (i.e., five of what were black dots in Figure 1). We do these five random draws 50 times (Figure 2 shows only one of the 50 times). The correlation coefficient metrics that result are $\rho = 0.70$; $\tau_{AP} = 0.68$; $\tau_{GAP} = 0.65$, averaging over 50 such random draws of five pairs to swap. In Figure 3, we randomly select system pairs from among the set of pairs that are not statistically significantly different from each other (i.e., five of what were white dots in Figure 1). We do this 50 times. The correlation coefficient metrics that result are $\rho = 0.99$; $\tau_{AP} = 0.92$; $\tau_{GAP} = 0.97$, averaging over 50 such random draws of pairs to swap. In general, we can see that swapping system pairs with large gaps yields lower correlation scores with any measure than swapping system pairs with small gaps. But the key observation to make is that, as proven in theorem 4, when the swapped system pairs have relative large gaps, $\tau_{GAP} = 0.65$ is lower than $\tau_{AP} = 0.68$; whereas when the swapped system pairs have relatively small gaps, $\tau_{GAP} = 0.97$ is larger than $\tau_{AP} = 0.92$.

4.2 Weights for System Pairs

Figures 4, 5 and 6 show the weight for each system pair in evaluating the correlation coefficient of two ranked lists under the measurement of ρ, τ_{AP} and τ_{GAP} respectively. The systems are also ranked by their ground truth MAP scores. So cell (1, 61) represents the weight of only swapping the systems at rank 1 and rank 61. In detail, taking τ_{GAP} in Figure 6 as example, the value of cell (1, 61) is calculated by: (1) produce the prediction list by only swapping the systems at rank 1 and 61 in ground truth list; (2) calculate the value of τ_{GAP} between the ground truth list and the prediction list; (3) fill the value of $(1 - \tau_{GAP})$ to cell (1, 61); (4) after filling in all the cells, rescale the matrix to (0, 1) and build the heatmap. Therefore, in general, darker cells represent system pairs with higher weight in evaluating the correlation coefficient.

Figure 4 shows the heatmap of system pair weights by using ρ. Since ρ is only sensitive to the score difference, not the ranks of items, we can observe that the value of ρ is dominated by the swapping pairs with large differences, probably composed of a system at very high rank and a system at very low rank. The system pairs along the diagonal with non-significant difference get lower weights as expected. However, the swap of top ranked systems also get lower weights because of their relatively small difference. The heatmap of τ_{AP} in figure 5 shows a progressively decreasing from the top-left corner of (1, 1) to the bottom-right corner of (61, 61) due to its sensitivity over ranks. However, comparing Figure 5 with Figure 1, we can see that the 37% non-significant system pairs still get non-ignorable weights in calculating τ_{AP}. Figure 6 shows the heatmap using τ_{GAP}. Since τ_{GAP} is sensitive to both top ranked systems and the gap between system pairs, we can also observe a decreasing of weights for the system pairs from top-left corner to the bottom-right corner. At the same time, we can see a expansion of the lower weight area (bright-ish area) along the diagonal line comparing with the heatmap of τ_{AP}. This is due to τ_{GAP}'s sensitivity to the system pairs with small difference. Overall, the value of τ_{GAP} is dominated by the top ranked system pairs with large difference.

5. CONCLUSION

In this paper, we proposed a new metric τ_{GAP} for evaluating the correlation coefficient between two ranked lists of scores. We have shown that τ_{GAP} is sensitive to both top ranked items and to swapped item pairs that exhibit larger differences. Through analysis, we have shown that τ_{GAP} compares favorably with both ρ and τ_{AP}, and using the TREC 5 Ad Hoc track as a case study we have illustrated that the swaps that τ_{GAP} is most sensitive to the ones we have argued we should care the most about. Although we have introduced τ_{GAP} in the context of system comparison, it is of course a general correlation coefficient for ranked lists of scores that could be applied in any case in which both a head-weighted and gap-sensitive measure would be useful. There are, however, some characteristics of τ_{GAP} that might be further improved upon. For example, it might be useful to completely discount differences in scores that are not statistically reliable indicators of real differences in system behavior. As Figure 1 illustrates, such cases can

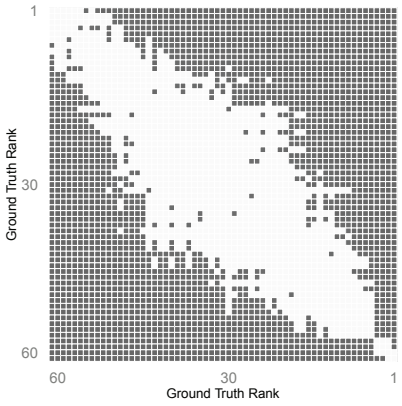

Figure 1: Significance of the MAP difference on TREC 5.

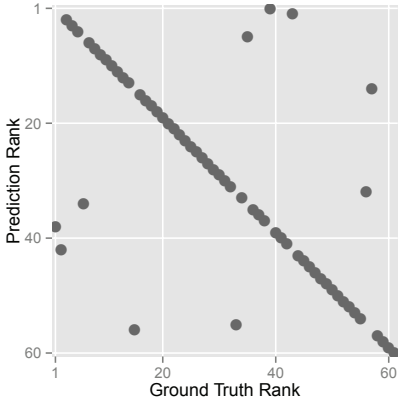

Figure 2: Significant difference between system pairs.

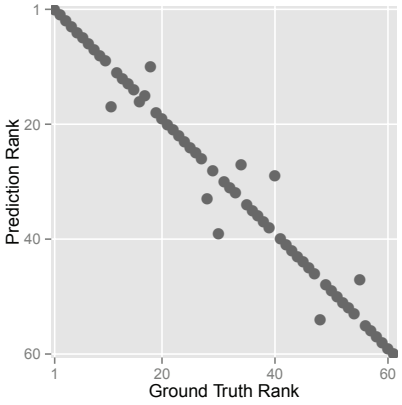

Figure 3: No significant difference between system pairs.

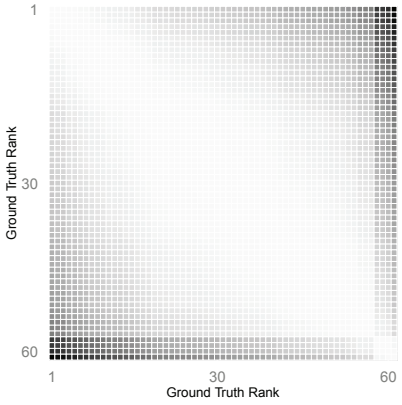

Figure 4: Gap-Sensitive: ρ

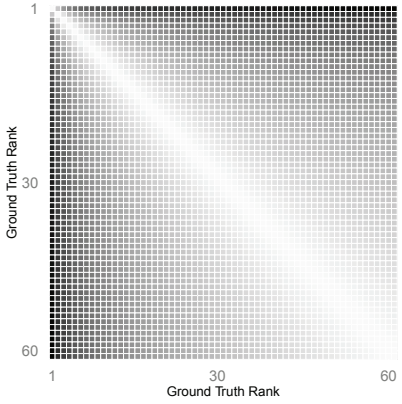

Figure 5: Head-Weighted: τ_{AP}

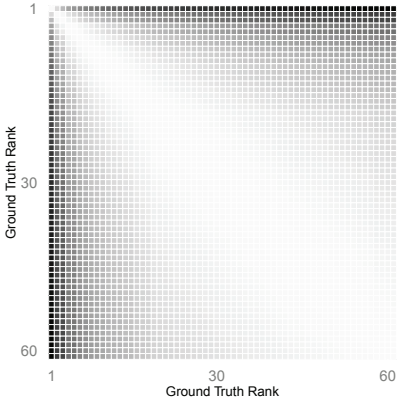

Figure 6: The proposed τ_{GAP}

be common. As another example, τ_{GAP} implicitly presumes that the scores are represented on a meaningful interval scale, meaning that (for example) a user would prefer a difference of 0.2 twice as much as they would prefer a difference of 0.1. User studies have shown that some current evaluation measures do not exhibit anywhere near this degree of correlation to extrinsic measures of success such as task completion rates [8]. Future extensions that more closely model extrinsic measures of satisfaction or success might therefore be useful. Nonetheless, we see the progression from τ to τ_{AP} and now to τ_{GAP} to be a useful one, and one that can perhaps serve as a basis for future extensions of these and other kinds.

6. ACKNOWLEDGEMENT

This work has been supported in part by NSF award 1065250. Opinions, findings, conclusions and recommendations are those of the authors and may not reflect NSF views.

References

[1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *SIGIR*, pages 33–40, 2000.

[2] N. Gao et al. Reducing reliance on relevance judgments for system comparison by using expectation-maximization. In *ECIR*, pages 1–12. 2014.

[3] K. S. Jones. Automatic indexing. *Journal of Documentation*, 30(4):393–432, 1974.

[4] M. G. Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

[5] M. G. Kendall. *Rank correlation methods.* Griffin, 1948.

[6] K. Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895.

[7] M. Smucker et al. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*, 2007.

[8] A. Turpin and W. R. Hersh. Why batch and user evaluations do not give the same results. In *SIGIR*, pages 225–231, 2001.

[9] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Process. Manage.*, 36(5):697–716, 2000.

[10] E. Yilmaz et al. A new rank correlation coefficient for information retrieval. In *SIGIR*, pages 587–594, 2008.

[11] G. U. Yule. *An introduction to the theory of statistics*. C. Griffin, limited, 1919.