Low-Cost Preference Judgment via Ties

Kai Hui[1,2(✉)] and Klaus Berberich[1,3]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
{khui,kberberi}@mpi-inf.mpg.de
[2] Saarbrücken Graduate School of Computer Science, Saarbrücken, Germany
[3] htw saar, Saarbrücken, Germany

Abstract. Preference judgment, as an alternative to graded judgment, leads to more accurate labels and avoids the need to define relevance levels. However, it also requires a larger number of judgments. Prior research has successfully reduced that number to $\mathcal{O}(N_d \log N_d)$ for N_d documents by assuming transitivity, which is still too expensive in practice. In this work, by analytically deriving the number of judgments and by empirically simulating the ground-truth ranking of documents from TREC Web Track, we demonstrate that the number of judgments can be dramatically reduced when allowing for ties.

1 Introduction

Offline evaluation in information retrieval heavily relies on manual judgments to generate a ground-truth ranking of documents in response to a query. There exist two approaches to collect judgments, namely, graded judgments, where documents are labeled independently with predefined grades, and preference judgments, where judges provide a relative ranking for a pair of documents. For instance, given a test query, there are two rivaling systems s_1 and s_2, whose search results in response to the test query are (d_3, d_1, d_2) and (d_5, d_4, d_2) respectively. To compare these two search results, manual judgments are collected. When collecting graded judgments, the five documents are assessed by judges independently and are assigned predefined grades, say $d_1 : 0$, $d_2 : 1$, $d_3 : 1$, $d_4 : 2$, $d_5 : 2$; when collecting preference judgments, pairwise preferences over document pairs are collected, say $d_5 \sim d_4$, $d_5 \succ d_3$, $d_5 \succ d_2$, $d_5 \succ d_1$, $d_4 \succ d_3$, $d_4 \succ d_2$, $d_4 \succ d_1$, $d_3 \sim d_2$, $d_3 \succ d_1$, $d_2 \succ d_1$. We use $\succ$, $\prec$ and $\sim$ to denote the "better than", "worse than", and "tied with" relationships. Ultimately, with both approaches, a ground-truth ranking of documents can be determined. In this example, a same ground-truth ranking of documents is derived from graded and preference judgments: $d_5 \sim d_4 \succ d_3 \sim d_2 \succ d_1$.

Preference judgments have been demonstrated as a better alternative to the widely used graded judgments. Compared with graded judgments, preference judgments lead to better inter-assessors agreement, less time consumption per judgment [3] and better judgment quality in terms of agreement to user clicks [5]. Radinsky and Ailon [7] pointed out that these advantages come from the pairwise nature of preference judgments, i.e., the documents in the pair can mutually

J.M. Jose et al. (Eds.): ECIR 2017, LNCS 10193, pp. 626–632, 2017.
DOI: 10.1007/978-3-319-56608-5_58

act as a "context", providing a reference for the judges. However, this pairwise nature also increases the number of judgments from $\mathcal{O}(N_d)$ to $\mathcal{O}(N_d^2)$ for N_d documents. Even after assuming transitivity, the number of judgments is still in $\mathcal{O}(N_d \log N_d)$ and hence much larger than the one from graded judgments, which is especially true for large N_d.

In this work, we highlight the ties in preference judgments, which have been introduced in existing works [5,8], but without noticing its potential in reducing the number of judgments. We assume transitivity among preference judgments as in [3,6,8], which might be over-optimistic in practice. We argue that, however, the collection of transitive judgments, and the design of judgment mechanisms that can tolerate intransitive judgments are orthogonal to this work. Moreover, the ultimate judgment cost should be the number of judgments times the cost per judgment, where a higher unit cost may lead to better transitivity. Instead we focus on demonstrating the potential of ties in reducing the number of judgments when transitivity is strictly observed. We investigate the number of judgments when allowing for ties analytically and empirically. In particular, we reexamine the number of preference judgments on N_d documents with established QuickSort-Judge mechanism [8]. Moreover, we empirically investigate the number of judgments when simulating the ground truth from Trec Web Track 2011–2014. To this end, we argue that the tie is a compromise between the number of judgments and the judgment granularity. It clusters documents into tie partitions, and reduces the ranking of documents to the ranking of tie partitions. We demonstrate that the average number of judgments is reduced to $\mathcal{O}(N_t \log N_d)$, where N_t is the number of tie partitions. In addition, when simulating the ground truth from Trec, compared with graded judgments, only 43% more judgments are required when allowing for ties, whereas 773% more judgments are required in strict preference judgments. To the best of our knowledge, this is the first work to investigate and confirm the importance of ties in reducing the number of judgments.

Organization. Section 2 recaps existing literature and puts our work in context. Section 3 analyzes how the ties can help to reduce the number of judgments analytically and empirically. Finally, in Sect. 4 we draw conclusions.

2 Related Work

Reduce the Number of Judgments. Assuming transitivity among preference judgments, the complexity is reduced from $\mathcal{O}(N_d^2)$ to $\mathcal{O}(N_d \log N_d)$ [1,3,8], by avoiding a full comparison among all document pairs. Beyond transitivity, several attempts to further bring down the number of judgments were made. Carterette et al. [3] proposed to remove 20% "Bad" judgments by assigning them as worse than others. Niu et al. [6] addressed the expensiveness by only determining a full order for top-k search results, reducing the complexity to $\mathcal{O}(N_d \log k)$. Actually, the documents labeled as "Bad" in [3] and the documents out of top-k in [6] can be regarded as special cases of tie partitions–a single tie partition with low relevance documents. However, we argue that the reduction of the number of

judgments is limited compared with a real tie option, which is especially true for the "Bad" judgments, given that the limited number of documents that are totally off-topic in practice. Moreover, the top-k ground-truth ranking from [6] is more suitable for learning to rank algorithms, and may lead to bias for evaluation purpose especially when smaller k is used. Other than that, no existing work has explicitly investigated the usage of ties in reducing the number of judgments.

Quick-Sort-Judge. In our empirical analysis, we employ the labeling mechanism Quick-Sort-Judge from [8], similar to a randomized *QuickSort* method. In Quick-Sort-Judge, during each iteration, a document is randomly chosen as a pivot document, denoted as d_p. Thereafter, all remaining documents are grouped into worse than ($\prec d_p$), better than ($\succ d_p$) or tied with ($\sim d_p$) per manual judgments. The mechanism terminates when all documents have been recursively sorted. Note that, within each iteration, the documents on different sides of the pivot document are not manually judged, instead preferences between such document pairs are inferred with transitivity.

3 Number of Judgments

In this section, we investigate the average number of judgments required by preference judgments with ties analytically and empirically.

3.1 Theoretical Analysis

We reexamine the expected number of preference judgments when allowing for ties based on Quick-Sort-Judge [8] as introduced in Sect. 2.

Notation. Given query q, we denote a set of documents as $\mathcal{D}$, and thus $N_d = |\mathcal{D}|$. Akin to the notation in [8], in the ground-truth ranking of documents on $\mathcal{D}$, documents that are mutually tied constitute N_t tie partitions, which are denoted as $t_1, t_2, \cdots, t_{N_t}$. Within individual tie partition t_i, documents are labeled with the same grade or are judged as mutually tied. For example, the ground-truth ranking of documents in the example from Sect. 1 can be represented as $t_1 \prec t_2 \prec t_3$, where $t_1 = \{d_1\}$, $t_2 = \{d_2, d_3\}$ and $t_3 = \{d_4, d_5\}$. Given tie partitions $t_i \prec t_j$, we use $\mathcal{D}_{ij}$ to denote documents which lie in between t_i and t_j in the ranking, namely, $\mathcal{D}_{ij} = \{d | t_i \prec d \prec t_j\}$. The set of tie partitions on $\mathcal{D}$ is denoted as $\mathcal{T}$. We introduce $\beta = \frac{N_d}{N_t}$, denoting the average number of documents per tie partition. Manual judgments can be categorized into two kinds: non-tie judgments, namely $\prec$and$\succ$, which sort tie partitions; and tie judgments, namely $\sim$, which cluster documents into tie partitions. Correspondingly, the total number of judgments, denoted as N_{jud}, can be split into the number of non-tie judgments, denoted as N_{ntie}, and the number of tie judgments, denoted as N_{tie}. And N_{ntie} can be further boiled down to judgments that determine relative order between a pair of tie partitions t_i and t_j, denoted as N_{ij}, namely, $N_{ntie} = \sum_{t_i, t_j \in \mathcal{T}} N_{ij}$.

Assumptions. As mentioned in Sect. 1, our analysis is based on transitivity assumption. The transitivity can be applied among tie partitions. For instance,

Table 1. The distribution and expectation of N_{ij}, namely, the number of judgments to determine the relative order of two tie partitions t_i and t_j.

Pivot document d_p	$t_i \prec d_p \prec t_j$	$d_p \in t_i$	$d_p \in t_j$
N_{ij}	0	$\lvert t_j \rvert$	$\lvert t_i \rvert$
$P(N_{ij})$	$\frac{\lvert \mathcal{D}_{ij} \rvert}{\lvert t_i \rvert + \lvert \mathcal{D}_{ij} \rvert + \lvert t_j \rvert}$	$\frac{\lvert t_i \rvert}{\lvert t_i \rvert + \lvert \mathcal{D}_{ij} \rvert + \lvert t_j \rvert}$	$\frac{\lvert t_j \rvert}{\lvert t_i \rvert + \lvert \mathcal{D}_{ij} \rvert + \lvert t_j \rvert}$
$E(N_{ij})$	$\frac{2\lvert t_i \rvert \lvert t_j \rvert}{\lvert t_i \rvert + \lvert \mathcal{D}_{ij} \rvert + \lvert t_j \rvert}$		

given t_i and t_j, by judging $d_k \in t_i$ and $d_l \in t_j$ as tied, one can get $t_i \sim t_j$ according to transitivity. In addition, we assume that $|t_\cdot| = \frac{N_d}{N_t} = \beta$, namely, tie partitions have the same size. Note that the size of different tie partitions is more skewed in practice, and this assumption is used to simplify Eq. 1.

Non-tie Judgments: Sort the Tie Partitions. For the non-tie judgments, the number of judgments is analyzed following the analysis for randomized *QuickSort* algorithm [4]. Akin to [4], conceptually, we index these tie partitions according to their ground-truth order, namely, $t_1 \prec t_2 \prec, \cdots, t_i \prec t_j, \cdots, t_{N_t}$. To approach this ground-truth order, one needs to determine relative order for each pair of tie partitions, say t_i and t_j. Therefore, one has to either select pivot document d_p from t_i or t_j, resulting in $|t_j|$ or $|t_i|$ judgments respectively, or select a pivot document d_p in between t_i and t_j, namely $d_p \in \mathcal{D}_{ij}$, leading to 0 judgments. In the former case, assuming $d_p \in t_i$, one needs to judge d_p relative to each document in t_j and make $|t_j|$ judgments. In the latter case, the relative order between t_i and t_j is inferred from the judgments between them and d_p, e.g., $t_i \prec d_p, t_j \succ d_p \implies t_i \prec t_j$. The distribution of the random variable N_{ij} is summarized in Table 1. And the expected total number of non-tie judgments $E(N_{ntie})$ can be computed as follows.

$$\begin{aligned} E(N_{ntie}) = E(\sum_{t_i,t_j \in \mathcal{T}} N_{ij}) &= \sum_{i=1}^{N_t-1} \sum_{j=i+1}^{N_t} E(N_{ij}) \\ &= \sum_{i=1}^{N_t-1} \sum_{j=i+1}^{N_t} \frac{2|t_i||t_j|}{|t_i| + |\mathcal{D}_{ij}| + |t_j|} \end{aligned} \tag{1}$$

Assuming that tie partitions have equal size, the complexity can be simplified as in Eq. 2, where $H_{N_t} = \sum_{k=1}^{N_t} \frac{1}{k}$ is the n_t-th harmonic number, which is in $\mathcal{O}(\log N_t)$ [4].

$$\begin{aligned} E(N_{ntie}) &= \sum_{i=1}^{N_t-1} \sum_{j=i+1}^{N_t} \frac{2\beta^2}{\beta(j-i+1)} \\ &= 2\beta \sum_{i=1}^{N_t-1} \sum_{k=2}^{N_t-i+1} \frac{1}{k} \\ &< 2\beta \sum_{i=1}^{N_t} H_{N_t} = 2\beta N_t H_{N_t} \end{aligned} \tag{2}$$

Tie Judgments: Generate Tie Partitions. When two documents are judged as tied, they are put into the same tie partition. For tie partition t_i, one needs to make $|t_i|$ tie judgments. Therefore, the total number of tie judgments is $E(N_{tie}) = \sum_{i=1}^{N_t} |t_i| = N_d$.

Total Number of Judgments. Henceforth, the expected total number of judgments equals the sum of the aforementioned two parts as in Eq. 3, which is in $\mathcal{O}(N_d \log N_t)$.

$$\begin{aligned} E(N_{jud}) &= E(N_{ntie}) + E(N_{tie}) \\ &< 2\beta N_t H_{N_t} + N_d \end{aligned} \tag{3}$$

3.2 Empirical Analysis

In this section, we empirically examine the number of judgments required in preference judgments to simulate the ground truth fromTrec.

Dataset. Our experiments are based on graded judgments from the 2011–2014 Trec Web Track[1] for adhoc task including 200 queries. The judgments contain at most six grades and one can sort them to establish a ground-truth ranking of documents.

Methods Under Comparison. We compare the number of judgments from three methods: graded judgments, preference judgments with ties and strict preference judgments. The number of judgments in graded judgments simply equals the number of documents. The preference judgments are simulated by randomly selecting document pairs with the established Quick-Sort-Judge [8] as introduced in Sect. 2. Thereafter, in preference judgments with ties, the judgments are simulated by comparing the ground-truth labels of two documents from Trec. For strict preference judgments, given that ties are not allowed, the relative order between documents with the same labels from Trec are further determined by their string identifiers, which are unique and fixed among random experiments. We report the average number of judgments from 1000 repetitions of Quick-Sort-Judge for both kinds of preference judgments.

Results. The results are summarized in Fig. 1. It can be seen that, the judgments from strict preferences are far more than the one when allowing for ties, namely, on average 500% more judgments are required. Compared with the number of judgments required by graded judgments, the numbers are 43% and 773% higher respectively when allowing and not allowing for ties.

3.3 Discussion

Results from Sects. 3.1 and 3.2 demonstrate that ties can dramatically reduce the number of judgments. Compared with strict preferences, ties actually produce coarser ground-truth rankings. This can be seen from the analytical results

[1] http://trec.nist.gov/tracks.html.

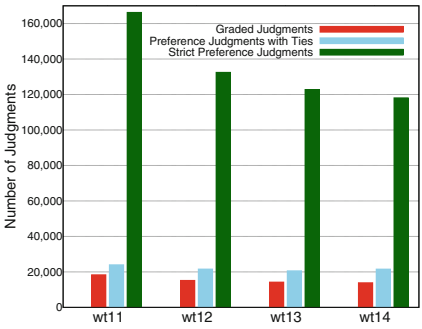

Fig. 1. The average number of judgments required by graded judgments and by preference judgments with/without ties on TREC Web Track. The x-axis is different years and y-axis represents the number of judgments. The averaged number of judgments from 1000 repetitions is reported as the actual number of judgments for both kinds of preference judgments.

$\mathcal{O}(N_d \log N_t)$ from Sect. 3.1: when $N_t = N_d$ ($\beta = 1$) it becomes strict preferences; and the number is reduced when $N_t < N_d$, where more documents are "squeezed" into a single tie partition. Meanwhile, the ground-truth ranking of documents is simplified to the ranking of tie partitions. In the example from Sect. 1, $d_2 \sim d_3$ and $d_4 \sim d_5$ are in the ground-truth ranking, meaning that the ground-truth relative rankings in between d_2 and d_3 and in between d_4 and d_5 are undetermined. In other words, the relative rankings between them are not considered in the evaluation as in [2]. Thus, the ties can be regarded as a compromise between the number of judgments and the judgment granularity.

Finally, we discuss whether there is potential to reduce the number of judgments with ties beyond QUICK-SORT-JUDGE. Similar to the strategy employed in [9], ideally, one can first make tie judgments to cluster documents, and thereafter make non-tie judgments to sort the tie partitions. By doing this, the number of tie judgments remains the same, namely N_d. Whereas for non-tie judgments, the number of judgments under $d_p \in t_i$ and $d_p \in t_j$ becomes 1 in Table 1, which means that one only needs to judge a pair of documents to determine the relative order of two established tie partitions. Accordingly, the number of judgments is reduced to $E(N_{jud}) = 2\sum_{i=1}^{N_t-1}\sum_{k=2}^{N_t-i+1}\frac{1}{k} + N_d < 2N_tH_{N_t} + N_d$, which is in $\mathcal{O}(2N_t \log N_t + N_d)$ and is close to linear when $N_t \ll N_d$.

4 Conclusion

In this work, we analytically derive and empirically simulate the number of judgments required in preference judgments. We demonstrate that the number of judgments can be reduced by simply allowing for ties, from $N_d \log N_d$ to $N_d \log N_t$. For future works, as discussed in Sect. 3.3, novel judgment mechanisms are desired to better utilize ties.

References

1. Ailon, N., Mohri, M.: An efficient reduction of ranking to classification. arXiv 2007 (2007)
2. Carterette, B., Bennett, P.N.: Evaluation measures for preference judgments. In: SIGIR 2008 (2008)
3. Carterette, B., Bennett, P.N., Chickering, D.M., Dumais, S.T.: Here or there: preference judgments for relevance. In: Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (eds.) ECIR 2008. LNCS, vol. 4956, pp. 16–27. Springer, Heidelberg (2008). doi:10.1007/978-3-540-78646-7_5
4. Cormen, T.H.: Introduction to Algorithms. MIT Press, Cambridge (2009)
5. Kazai, G., Yilmaz, E., Craswell, N., Tahaghoghi, S.M.: User intent and assessor disagreement in web search evaluation. In: CIKM 2013 (2013)
6. Niu, S., Guo, J., Lan, Y., Cheng, X.: Top-k learning to rank: labeling, ranking and evaluation. In: SIGIR 2012 (2012)
7. Radinsky, K., Ailon, N.: Ranking from pairs and triplets: information quality, evaluation methods and query complexity. In: WSDM 2011 (2011)
8. Song, R., Guo, Q., Zhang, R., Xin, G., Wen, J.-R., Yu, Y., Hon, H.-W.: Select-the-best-ones: a new way to judge relative relevance. Inf. Process. Manag. **47**(1), 37–52 (2011)
9. Wang, J., Li, G., Kraska, T., Franklin, M.J., Feng, J.: Leveraging transitive relations for crowdsourced joins. In: SIGMOD 2013 (2013)