How Precise Does Document Scoring Need To Be?

Ziying Yang

Slide 0

(Greeting)

I am Ziying Yang. Today I am going to present my work cooperated with Alistair Moffat and Andrew Turpin from the University of Melbourne, about how precise does document scoring need to be.

Slide 1

When we want to answer the question, given two IR systems, A and B, which one can do better than the other, we need a process of evaluation for both systems. Usually we use a Batch Evaluation Technique.

With this technique, for each system, and for each of all given topics, (1) we measure how documents are relative to the topic and assign each of them a similarity score using some ranking functions such as BM25. (2) According to these document scores, the system will generate a ranking list, that we can call a run, containing documents relative to this topic. The documents in the list are in decreasing score order. (3) Then, we select an effectiveness metric, such as Average Precision, to evaluate this run based on a set of relevance judgments and give the run a score.

We repeat these steps for every given topic, and then aggregate run scores for all the given topics into a single system score, which can be seemed as the overall performance quality of the system. Finally, we can compare system A and B using their system scores.

Slide 2

It need to be noticed that the effectiveness metrics only care about ranks. They only concern about in this ranking list, at each rank, how many relevance gains can be obtained by users. They do not care about the document similarity scores at all.

But this is a problem for us. If documents are assigned the same scores for the similarity, we call them ties, they may be two, three or even more than ten documents, how can we order them in the ranking list?

Please notice that the ties that I mention in this speech only refer to the documents which have the same similarity scores. We do not concern about the tied run or tied system here.

As we know, the relevance gain of these tied documents can vary greatly, some of them are relevant but some of them are not, for example. So if we rank them in different ways, we may receive different system scores given by metrics, and it may influence the final outcome of the experiment.

Slide 3

Let's look at an example of this. Here is a ranking list for a topic. In this table, the header is rank. Document at each rank is named as a letter. The third row is the gain of each document. Here we use the binary relevance judgment in the example, so 0 represents that this document is not relevant and 1 represents that it is relevant to the topic. And the next row is the document score. As you can see, they are in decreasing order and the scores of H, A and C are the same, 9.3, so they are ties. And so as document M and S. We divide this list into groups, shown as the last row. Documents in each group have the same scores. (D in group 1, ...). We denote b_g as the beginning rank of the gth group...

Slide 4

For the ties in the ranking list, we have some possible methods to deal with them when we evaluate the run.

First option, use the run order. The system has to pick an order for the tied documents by some mechanisms because the documents in the ranking list must be linearized for sequential presentation. So we just use this order.

The second option is to re-order the tied documents in each group using the fixed ordering criterion. For example, we can sort them by document ID and that is how `trec_eval` program does. In the group 2, documents will be re-order as: A, C and then H.

The third option is to find out both the optimistic and pessimistic limits. As we already have relevance judgments, for the ties in each group, we can put the relevant documents to the top of the group and non-relevant documents to the bottom. In this example, in the group 2, A and C will be put before the H because the gains of A and C are higher than the gain of H. In group 3, M and S will be switched as well. Then we compute the metric score in the usual way so we can find the best score that this run can get. Similarly, we can get the pessimistic lower score bound. Using this strategy, a score range instead of a single score value will be returned.

While the best-case and worse-case bounds can be informative, computing the average, or expected metric score across all possible permutations of documents within each of the tied score groups provides a useful balance. Using this strategy, every permutation of documents in each group is assumed to be equally. So we compute the metric scores for all permutation and take the average.

Slide 5

So in our first experiment, we tried to explore the role of ties in TREC evaluation, how many tied scores existed and how many systems, runs and documents were affected.

We used the TREC7 as the dataset. There are 103 submitted systems for 50 topics and (4 million 900 thousands and 42) documents in total. For each submitted run, we re-sorted it using decreasing score as a primary key, and increasing rank as a secondary key. It guaranteed that there were no score-based out-of-order items.

Then we counted how many tied scores existed at the levels of system, run and document. As you can see in the first row of this table, 95.2% of systems, 91% of runs, and 14% of documents were affected by the tied scores. That is, 14% of the documents in the runs have the same score as their previous documents in that run, which provides the motivation for our work here.

More than that we found that 1.4% of documents sorted by scores have ranks that indicate the opposite ordering, with 7 of the 103 systems affected. We can only assumed that it was caused by

programming errors when those research groups created the runs.

Slide 6

So when we examine the effect of ties on AP scores for systems, we used `trec_eval` to compute the mean AP scores over 50 topics for all the systems, and select only the top 80 systems for our next experiments to ensure that the results were not affected by mistakes or misunderstandings of the TREC7 participants.

For the selected systems, we compute three more AP scores for each of them. First, we compute the average AP scores across permutations and then we take the mean across topics. Then re-sorted the documents according to the relevance judgments in each tied score group and so we can compute the worst and the best mean AP scores across topics.

Slide 7

In this figure, we plot the results of this experiment.

There are totally 80 systems, each is plotted as a segment in this graph. For each system, the color represents how many ties were generated in this system: red means more than one thousand, yellow means between 501 and 1000, purple means between 1 and 500, and green means no ties were generated.

The x-axis is the absolute value of AP score. So the systems plotted at the most right hand side are the systems performed the best. The right and the left ends of the segment reflect the AP scores generated by the optimistic and pessimistic orderings for each of the tied groups, that is the upper and lower bounds of the AP score range of this system. The circle represent the AP score calculated by the `trec_eval`. And the triangle is the average across permutation score.

The y-axis, in logarithm scale, is AP score difference, also width of each segment. So the height of the segment is computed by the difference between the best and the wost AP score limits. So in the figure, the higher up the axis the system is plotted, the wider score range the system has, and of course their scores have greater uncertainty because of the tied scores. For the segments at the bottom, most of them do have ties but they do not result in any appreciable score range.

The circle, that is the `trec_eval` AP score, is just one within the score range, but usually it is not too far away from the average overall, the triangle. But for some systems, their score ranges are wide and overlap other systems. As you can see at the right top corner of this graph, there is a system which is ordered as the third best system by `trec_eval`. But the ties may have affected the relative ordering of the top few systems. If we randomly choose orderings for ties in this system, this system may have chance to be ranked at the first, second, third or the fourth in the competition. So in this round, it shows that ties do have potential to be disruptive to TREC evaluation.

We carried out the same analysis on some other TREC rounds. We found similar ties rates and also some systems with wide potential score ranges. But for those further years' rounds, we did not find the ordering of the top systems might have been affected by the tie-breaking rule employed.

Slide 8

We found that the ties were mostly caused by score rounding. But sometimes it is expensive and not necessary to keep scores with high precision even it will cause ties.

So in our next experiment, we aimed to find which level of score rounding can be tolerated without greatly affecting system compares. The finding will provide faster search by reducing the

score precision but keep the system compares safe from the generated ties.

Slide 9

We grouped documents into bands and assigned them the same scores. In other words, we generate ties deliberately in order to find how they affect the retrieval quality.

In the notations, g is the band serial number, should be an integer greater than or equal to 1. In the gth band, ranks are from b_g to e_g. For the first band, b_1 should be 1 of course. And then the beginning rank of the second band, b_2, should be the rounded up value of $\rho \times b_1$. And similarly, b_{g+1} should be the rounded up value of $\rho \times b_g$.

For example, if $\rho = 2$, b_2 should be equal to $2 \times b_1$, that is 2. And then b_3 is $2 \times b_2$, so it is 4.

And we can easily find the ending rank e_g of each band g, that is previous rank of the beginning rank of next band $g + 1$, that is $e_g = b_{g+1} - 1$.

In the following examples, we can see that the widths of bands are geometrical sequence and can be controlled by the value of ρ. If ρ is 1.62, the golden ratio, the first band only contains 1 document, the second band contains the following 2 documents, the third ...

Slide 10

In the experiment, each of the 80 systems generated a run for each of the 50 topics, so totally we have 400 runs. For each of them, we computed its effectiveness score of metric Reciprocal Rank, Rank Biased Precision with p value of 0.5 and 0.85, and Average Precision using the original run. Then apply the grouping rule described in the previous slide to the ranking list, with a set of ρ. And compute the effectiveness scores of the same 4 metrics using the banded run. Finally we plotted the differences between the original run score and the banded run score.

Slide 11

This is the figure of variation in AP scores. The x-axis is the value of ρ. The y-axis is the difference between the original run score and the banded run score. There are 400 runs in total so 400 score differences were used to plot the box graph in each column. The colorful crosses are outlier outside the 1.5 times of the inter-quartile range from the box limits.

As you can see, when $\rho = 1$, no ties was deliberately generated. The original run and the banded run are exactly the same. So all the score differences are zero. When the ρ grows larger, each time grows by 0.1, until $\rho = 2$, the score differences spread wider. And it is easy to find that, on average, the scores of banded runs are smaller than the scores of the original runs because of the ties.

Slide 12

You may find that the absolute values of these differences are relatively small, because AP is a deep metric and the AP scores themselves were small as well. If we look at the differences of RBP (p=0.85) scores, the limits of the boxes become wider which indicates the score differences are greater. We still have the same pattern that as the ρ grows, the score differences increase and the ties affect the evaluation more.

Slide 13

With RBP0.85, users are assumed to have 85% chance to keep checking the document at the next rank. If we change the value of p from 0.85 to 0.5 for RBP, which means that users now are assumed to have only 50% chance to move on. So RBP0.5 is shallower than PBP0.85 because users are expected to stop checking the ranking list earlier, and so which treats the change of the top part more important.

In this figure for differences of RBP0.5 scores, the box ranges are wider than the box ranges shown before. RBP0.5 is the shallowest metric so it suffers the most from the score grouping process.

And you may find that, the score differences have a big jump when ρ is larger than 1.5. This because when ρ is larger than 1.5, the ranks 2 and 3 start to be placed in the same group. As the first group always contains a single document when ρ is less than 2.0, ties in the second group will bring large differences to the scores.

Slide 14

We further explored whether the score differences in these graphs can be regarded as being significant. The original scores were multiplied by 0.97, and then compared with the banded scores, using a one-tail paired t–test.

If the generated p value for a system was less than or equal to 0.05, means that the grouping process degraded this system score by 3% or less. The closer the count of such systems is to 80 (remember we have 80 systems in total), then with that value of ρ, we have more confidence that the grouping process will not degrade 3% of the original scores.

The results are shown in the table. Reciprocal Rank is almost unaffected by the grouping process, no matter what value the ρ is. But for the rest of three, as the ρ increases, the chance that the scores change increases as well, and that is why less systems is counted as not affected. And we found that the number of systems decrease faster for the shallower metrics.

Slide 15

Effectiveness measurements are also used to compare systems in pairwise manner. In our last experiment, we explored how the score rounding affects the metrics to differentiate systems.

In the t–test, for each of the 80 system, we compare it with the other 79 systems using their run scores across 50 topics given by metrics. So totally $80 \times 79/2$ system pairs were generated. The null hypothesis is that in each pair, the compared two systems are the same. The t–test will give a p value for each system pair. And we plotted them in figures.

Show Graphs

Slide 16 We have explore the impact of similarity score ties on the IR system evaluation, measured using binary relevance judgments and metrics RR, RBP and AP.

We have showed that ties do have the potential to affect system comparisons. With the currently employed ties-breaking rules, a small number of TREC systems did generate runs with ties and system scores with wide ranges. But fortunately, the overall conclusions of those TREC rounds were unlikely to be compromised.

Then we demonstrated that even deliberately group documents as ties in runs, the ability of metrics to compare systems were almost not affected, which means that it is not necessary to keep similarity scores with high accuracy to some extends, even some ties will be generated. It is fine

because it will nearly not affect the system comparisons. Allowing the score rounding can help to increase the search speed and reduce the space used.