Are Evaluation Metrics Identical With Binary Judgements?

Milad Shokouhi Emine Yilmaz Nick Craswell Stephen Robertson
Microsoft Research Cambridge
{milads,eminey,nickcr,ser}@microsoft.com

ABSTRACT

Many information retrieval (IR) metrics are top-heavy, and some even have parameters for adjusting their discount curve. By choosing the right metric and parameters, the experimenter can arrive at a discount curve that is appropriate for their setting. However, in many cases changing the discount curve may not change the outcome of an experiment. This poster considers query-level directional agreement between DCG, AP, P@10, RBP($p = 0.5$) and RBP($p = 0.8$), in the case of binary relevance judgments. Results show that directional disagreements are rare, for both top-10 and top-1000 rankings. In many cases we considered, a change of discount is likely to have no effect on experimental outcomes.

1. INTRODUCTION

In the field of information retrieval, many different evaluation metrics have been proposed and used. Each of these metrics is believed to evaluate different aspects of retrieval effectiveness. Hence, much research has been devoted to identifying what constitutes a good metric and which metric to use for evaluation [1, 2].

Since users care more about the top end of the ranking, most evaluation metrics employ a discount function that aims at modelling how much users care about each ranking. The discount functions used by some metrics are parametric, and different methods have been used to learn the optimal values of these parameters [5, 6].

In this poster, we consider four of the most commonly used metrics in IR, precision at 10 (P@10), DCG [3], rank biased precision (RBP) [4], and average precision (AP). When the binary versions of these metrics are considered, the difference between these metrics is the discount function.

Precision at 10 (P@10), for example, assumes that users equally care for the top 10 documents and completely ignore the rest of the ranking.

Even though the discount function used in DCG [3] is not completely specified, most commonly used discounts are the $\frac{1}{\log_b(r+1)}$ (b specified depending on the persistence of the user) and the Zipfian $1/r$ discount, where r is the rank at which document is retrieved.

RBP assumes that the users scan the ranked list of documents from top to bottom and at each step may continue scanning the ranked list with some probability p or stop with probability $1-p$. Hence, the discount function used by RBP follows a geometric distribution.

SIGIR Workshop on the Future of IR Evaluation, July 23, 2009, Boston.

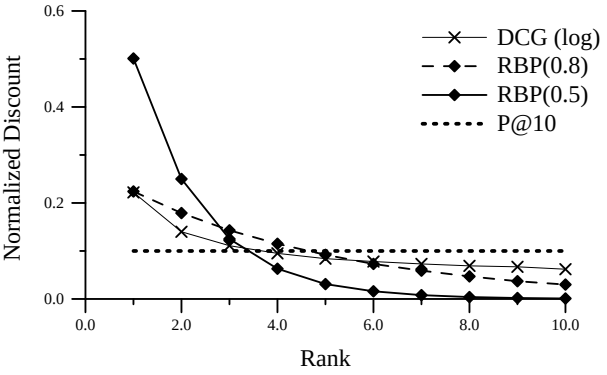

Figure 1: Normalized discount functions for different evaluation metrics. The discount function for AP is adaptive and not shown.

Average precision is defined as the average of the precisions at relevant documents. Therefore, discount function used by AP is adaptive; i.e., the discount of a document retrieved at rank r depends on the relevance of documents retrieved above rank r.

Figure 1 depicts the discount functions of different evaluation metrics. The numbers are normalized so that the area under each curve is equal to one. That is, each data point represents the importance of each rank according to a discount function. Even though these metrics seem quite different when the discount function is considered, what is important for evaluation purposes is whether they agree with each other on the relative quality of two different ranked lists. In this poster, we focus on the case of binary relevance judgments and we analyze whether the difference in the discount functions leads to different conclusions on the relative quality of rankings. In particular, we show that especially when real rankings are considered, most metrics agree on what is a better ranking. We conclude that using different discount functions (i.e., different evaluation metrics) actually leads to similar outcomes when judgments are binary.

2. EXPERIMENTS AND ANALYSIS

We measure the *agreement* rates of different metrics, by comparing their pair-wise preferences for various pairs of rankings. For a given pair of metrics M_a and M_b, and a given pair of ranked lists l_i and l_j, the metrics are in *agreement* if they both prefer the same list. That is, if $M_a(l_i) > M_a(l_j)$,

Table 1: The agreement rate between different metrics over several ranked lists (with different distribution of relevant and nonrelevant documents). In the *uniform* experiments, all ranked lists are considered equally likely. In the *sampled* experiments, the likelihood of ranked lists are approximated by using the previous TREC runs. Parameter N denotes the size of the ranked lists, and Δ is the *Fuzziness* value.

Metric Pairs	uniform	sampled	uniform	sampled	uniform	sampled	uniform	sampled
	$N = 10, \Delta = 0$		$N = 1000, \Delta = 0$		$N = 10, \Delta = 0.01$		$N = 1000, \Delta = 0.01$	
DCG/AP	0.95	0.98	0.90	0.92	0.96	0.99	0.99	0.99
DCG/P@10	0.75	0.92	0.54	0.70	0.93	0.97	0.76	0.81
DCG/RBP(0.5)	0.83	0.93	0.61	0.71	0.84	0.94	0.69	0.76
DCG/RBP(0.8)	0.94	0.98	0.64	0.74	0.96	0.98	0.72	0.78
P@10/AP	0.76	0.92	0.51	0.76	0.94	0.98	0.73	0.91
RBP(0.5)/AP	0.82	0.93	0.56	0.76	0.83	0.94	0.63	0.87
RBP(0.5)/P@10	0.60	0.86	0.59	0.81	0.77	0.92	0.76	0.90
RBP(0.8)/AP	0.94	0.98	0.59	0.80	0.96	0.98	0.67	0.89
RBP(0.8)/P@10	0.72	0.90	0.71	0.87	0.90	0.96	0.88	0.96
RBP(0.8)/RBP(0.5)	0.85	0.94	0.85	0.93	0.87	0.95	0.86	0.93

then $M_b(l_i) > M_b(l_j)$ and vice versa. In our experiments, we compare AP, DCG (logarithmic discount), P@10 and two variants of RBP with $p \in \{0.5, 0.8\}$. We use binary judgements for relevance, and consider the top-N documents in rankings to measure the agreement rates ($N \in \{10, 1000\}$).

The first five columns in Table 1, include the pairs of metrics, and their agreement rates for short ($N = 10$), and long ($N = 1000$) ranked lists. For short rankings ($N = 10$), there is a total possible of $\binom{1024}{2}$ ranking pairs that can be generated by varying the number of relevant documents in the top N. For each of these possible permutations, we calculate the value of each metric on both lists, and measure the ratio of *inter-metric* agreement accordingly. The agreement ratios computed this way, assume *uniform* likelihood for each pair of ranked lists. However, IR systems are biased towards returning more relevant documents on top of the ranked lists. Therefore, we also report the *sampled* version of agreement rates, by approximating the likelihood of each ranking according to previous TREC runs.[1] For long ranked lists ($N = 1000$), it is not feasible to try all the possible $\binom{2^{1000}}{2}$ permutations. Therefore, we generated about 5×10^7 random ranking pairs, where the probability of visiting a relevant document at each position is always 0.5.

The numbers in Table 1 suggest strong agreement rates between all the tested metrics for $N = 10$. In general, P@10 has the lowest agreement with the other metrics, which is not surprising given its shallow cutoff. For $N = 1000$, P@10 shows higher agreement rates with the two variants of RBP. This can be explained by aggressive discount function of RBP (Figure 1) that does not noticeably reward relevant documents at lower ranks. Furthermore, the agreement rates between the sampled lists are consistently higher than the uniform sample case. This shows that when metrics disagree, the disagreement is usually between the lists that are unlikely to appear in practice, and metrics mostly agree on the relative quality of reasonable ranked lists.

Fuzziness value (Δ). Buckley and Voorhees [2], defined the *Fuzziness value*, as the "the percentage difference between scores such that if the difference is smaller than the fuzziness value the two scores are deemed equivalent". The last four columns in Table 1 include the results for $\Delta = 0.01$. Here, the metrics M_1 and M_2 are in *disagreement* for a ranking pair l_i, l_j, *iff* (a) they prefer opposite rankings, and (b) $|M_a(l_i) - M_a(l_j)| > \Delta$, and $|M_b(l_i) - M_b(l_j)| > \Delta$. As was expected, employing a fuzziness threshold consistently boosts the agreement rates across all experiments.[2]

[1] We employed all the runs submitted to TREC7 and TREC8 ad hoc tracks. In total, there were 232 systems, each returned rankings for 50 queries.

3. CONCLUSIONS

We compared four of most commonly used evaluation metrics in information retrieval over millions of pairs of ranked lists. When all lists are considered equally likely, the metrics may look different than each other. However, in reality, not all lists are equally likely. In most cases, the probability of relevance decreases by rank. In order to identify whether metrics are different when reasonable ranked lists are considered, we used TREC runs to approximate the likelihood of each ranked list. When such a background distribution is employed, metrics seem highly correlated with each other, substantially more than uniform scenario. The agreement increases further by considering even small fuzziness intervals (e.g. $\Delta = 0.01$), to the extent that many metrics become almost identical (e.g. AP versus DCG). This suggests that most metrics agree on reasonable lists, and the most disagreements between metrics are only on the lists that are very unlikely to be real search results.

4. REFERENCES

[1] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proceedings of the ACM SIGIR conference*, pages 27–34, Salvador, Brazil, 2005.

[2] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the ACM SIGIR conference*, pages 33–40, Athens, Greece, 2000.

[3] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[4] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27(1):1–27, 2008.

[5] L. A. Park and Y. Zhang. On the distribution of user persistence for rank-biased precision. In *Proceedings of the Twelfth Australian Document Computing Symposium*, pages 17–24, Melbourne, Australia, 2007. School of CS and IT, RMIT University.

[6] K. Zhou, H. Zha, G.-R. Xue, and Y. Yu. Learning the gain values and discount factors of dcg. In *Proceedings of the Workshop on Beyond Binary Relevance: Preferences, Diversity, and Set-Level Judgments.*, Singapore, 2008.

[2] Increasing the value of Δ leads to further increase in agreement ratios. For example, when $\Delta = 0.05$ and $N = 10$, the agreement is always above 95%, except for RBP(0.5)/P@10 where it is 92%.