Features of Disagreement Between Retrieval Effectiveness Measures

Timothy Jones
RMIT University
timothy.jones@rmit.edu.au

Paul Thomas
CSIRO
paul.thomas@csiro.au

Falk Scholer
RMIT University
falk.scholer@rmit.edu.au

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

ABSTRACT

Many IR effectiveness measures are motivated from intuition, theory, or user studies. In general, most effectiveness measures are well correlated with each other. But, what about where they don't correlate? Which rankings cause measures to disagree? Are these rankings predictable for particular pairs of measures? In this work, we examine how and where metrics disagree, and identify differences that should be considered when selecting metrics for use in evaluating retrieval systems.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*

Keywords

IR evaluation, effectiveness measures, binary relevance

1. INTRODUCTION

Since information retrieval systems were first evaluated, system effectiveness measures have been a topic of discussion, study and controversy. When a new evaluation metric is introduced–or existing measures are criticised or praised–the discussion is usually motivated either by theoretical concerns with earlier metrics [5, 8, 9], or by user studies [1]. Sometimes, both approaches are combined [4].

Most IR effectiveness measures assume that the quality of results returned by a search engine can be calculated as a function of the gain vector inferred by a ranked list–sometimes also including additional knowledge such as the total number of relevant documents available [2]. In this paper we use:

- k as the depth to which a particular effectiveness metric has been evaluated.
- R as the total number of relevant documents available.

In the case of binary relevance (where documents are assumed to either be relevant, or not), the gain vector can be represented as a bit string. In simple cases, it is easy to examine these bitstrings by eye and reason about which list should be preferred. For example, where $k = 3$, $\{1, 0, 0\}$ is almost always better than $\{0, 0, 1\}$.

However, some cases become more contentious. Consider two vectors with $k = 10$ and $R = 10$:

$$A = \{1100010000\}$$
$$B = \{1000101100\}$$

If asked which list is better, most researchers would say "it depends on the task". However, it's not always immediately clear whether particular effectiveness metrics would agree on which list is better. In this particular example, AP prefers ranking A over B, while DCG prefers ranking B over A. Put another way, it's not always clear which metric prefers which type of task, or whether particular metrics are consistent in their preferences. In this work, we perform an exhaustive search of the possible binary relevance vectors where $k = 10$, and investigate where the disagreement between metrics lies.

2. RELATED WORK

It's worth noting that many metric descriptions don't completely specify the details of implementation. For example, an implementation of DCG requires a selection of both a gain and a discount function. Similarly, an implementation of RBP requires a selection of the p parameter [9]. Kanoulas and Aslam examine several different possible choices for the gain and discount functions in NDCG [6]. They used a Generalisability Theory approach to find choices that produced a stable ranking of systems. In this environment, they show that the optimal discount function is less steep than previously thought, and also that the optimal gain function gives nearly equal weight to relevant and highly relevant documents.

Although producing a single score for a retrieved list is convenient, a single score is not a particularly effective way of capturing system performance [7]. One illustration of this is

SIGIR '15, August 09–13, 2015, Santiago, Chile
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3621-5/15/08 ... $15.00.
http://dx.doi.org/10.1145/2766462.2767824.

Metric	Bounced	Monoton.	Converg.	Top-wgt	Localis.	Complete	Realis.
DCG	No	Yes	Yes	Yes	Yes	Yes	No
SP	No	Yes	Yes	Yes	Yes	Yes	No
RPrec	Yes	No	No	No	No	No	Yes
SN-DCG	Yes	No	No	Yes	Yes	No	Yes
SN-AP	Yes	No	No	Yes	Yes	No	Yes
Precision	Yes	No	Yes	No	Yes	Yes	No
NDCG	Yes	No	Yes	Yes	No	No	Yes
SDCG	Yes	No	Yes	Yes	Yes	Yes	No
RR/ERR	Yes	Yes	No	No	Yes	Yes	Yes
Recall	Yes	Yes	Yes	No	No	No	No
AP	Yes	Yes	Yes	Yes	No	No	No
RBP	Yes	Yes	Yes	Yes	Yes	Yes	No

Table 1: Properties of effectiveness metrics used (table reproduced and slightly modified from Moffat [8]. See original paper for footnotes and full discussion). All metrics are assumed to be an @k version.

the idea of 1-equivalence [10], which is the set of binary result vectors that receive the same score as a single document. The authors note that many 1-equivalent sets are contentious or at least counter-intuitive.

An interesting approach to measuring the quality of an effectiveness measure is to use the Maximum Entropy Method to infer the probabilities that each rank contains a relevant document, given the effectiveness score [2]. Once this probability vector is computed from the effectiveness score, a predicted precision–recall curve can be produced. The error between this and the actual precision–recall curve can be used to determine the quality of the metric. This approach asks the question “how good is the effectiveness score at inferring the original ranked list?”.

Motivated by the wide variety of effectiveness metrics available, Moffat [8] introduces seven properties for describing and comparing metrics. They are:

- Boundedness: Whether the range of scores a metric can produce is bounded or not. For example, NDCG is bounded in the range 0–1, while DCG is unbounded in the upper range.

- Monotonicity: If extra documents are appended to the tail end of a ranking, the new score produced by a monotonic ranking is never less than the previous score. For example, Recall-at-k and DCG are monotonic, while AP is not.

- Convergence: If a document within the top k is swapped with a more relevant document outside the top k, then the score always increases if the metric is convergent. For example, DCG is convergent, while RR is not.

- Top-weighted: If a document within the top k is swapped with a more relevant document also within the top k but lower down the ranking, then the score always increases if the metric is top-weighted. For example, AP is top-weighted, while Precision-at-k is not.

- Localisation: If a score at depth k can be produced using only the information about the documents down to depth k, then a metric is localised. For example, RR is localised, while NDCG and AP both require additional information about the relevant documents that did not make it in to the ranking.

- Completeness: If a score can be produced when a query returns no relevant documents, then the metric is complete. For example, DCG is complete, while NDCG and AP both produce a division by zero when there are no relevant documents.

- Realisability: If a collection has any relevant documents, then a metric is realisable if it is possible to achieve the maximum value for that metric. For example, RR is realisable, but RBP is not.

Moffat notes that it is impossible for a metric to satisfy all seven properties, as some property combinations exclude others (for example, a metric that is monotonic and convergent cannot also be realisable if $k < R$).

3. METRIC DISAGREEMENT

Many previous comparisons between effectiveness metrics have looked at correlation between metrics. In this work, we are interested in the specific cases where metrics do not agree. That is, the cases where we have two binary relevance vectors A and B where–for example–AP says that A is the most effective result list, while DCG says that B is the most effective result list. In this work, we only focus on rankings where $k \leq 10$, since many of the metrics above include a user model, and it is common for users to only examine the first page of results in a web setting.

3.1 Method

We generate all possible combinations for binary relevance rankings where $1 \leq k \leq 10$ and $1 \leq R \leq 10$. This yields a total of 61,430 possible ranked lists. Since we are interested in the disagreement between metrics on specific (potentially) real queries, we only consider pairs of lists for comparison where both lists A and B have equal values of k and R. There are 32,062,341 such pairs. Each ranked list is evaluated for each metric, and for each pair of ranked lists and pair of metrics, agreement or disagreement is recorded. All metrics were computed using 64-bit floating point numbers, and equality was checked using an epsilon of 2^{-53}.

3.2 Metrics

We use the same metrics as Moffat [8], with the exception of HIT, which would receive the same score for every ranked list described above.

	DCG	RR	P@k	SDCG	NDCG	RBP85	RBP95	RBP50	ERR	R-Prec	SP	R@k	SN-AP	SNDCG
AP	1.32	30.10	20.62	1.32	1.32	2.13	3.51	8.67	30.10	20.62	0.01	20.62	31.20	31.68
DCG		29.38	20.80	0.00	0.00	2.83	3.91	8.06	29.38	20.80	1.31	20.80	31.11	31.07
RR			40.50	29.38	29.38	31.55	33.06	23.82	0.00	40.50	30.09	40.50	40.95	40.62
P@k				20.80	20.80	19.45	17.78	28.81	40.50	0.00	20.62	0.00	51.64	51.87
SDCG					0.00	2.83	3.91	8.06	29.38	20.80	1.31	20.80	31.11	31.07
NDCG						2.83	3.91	8.06	29.38	20.80	1.31	20.80	31.11	31.07
RBP85							1.85	10.15	31.55	19.45	2.12	19.45	32.68	33.15
RBP95								11.97	33.06	17.78	3.50	17.78	34.52	34.98
RBP50									23.82	28.81	8.67	28.81	26.08	24.80
ERR										40.50	30.09	40.50	40.95	40.62
R-Prec											20.62	0.00	51.64	51.87
SP												20.62	31.20	31.67
R@k													51.64	51.87
SN-AP														1.88

Table 2: Disagreement between metrics (%, lower is better).

Property	Has-prop	No-prop	Cross-prop
Bounded	24.38	1.31	15.43
Monoton.	22.25	26.11	20.68
Converg.	**18.75**	29.51	24.35
Top-wgt	18.77	26.46	**25.47**
Localis.	19.64	24.33	24.63
Complete	19.41	25.39	24.33
Realis.	20.01	**19.24**	24.69

Table 3: Disagreement between metrics with particular properties (%, lower is better). Bolded groups show significance.

Metric combination	Recall	First-rel	First-irrel
AP	52.41	4.07	65.53
DCG	22.10	71.63	14.34
DCG	33.24	62.38	11.27
RBP85	37.71	11.27	62.38
RBP85	0.00	75.00	1.38
RBP95	89.96	1.38	75.00
RBP50	0.00	70.24	0.00
RBP95	89.89	0.00	70.24

Table 4: Features of result lists where metric pairs disagree.

The metrics used are DCG, SP, R-Precision, SN-DCG [8], SN-AP [8], Prec@k, NDCG [6], SDCG [8], RR, Recall@k, AP [3] and RBP [9]. For RBP, 3 different values of p were used: 0.5, 0.85 and 0.95. As recommended by the authors [9], the lower bound of the score was taken to be the RBP score for this experiment. ERR was also implemented, but since it receives the same score as RR in binary relevance, the results are identical to RR [4]. See Table 1 for the breakdown of properties of each measure. Where choices for implementation were available, we implement the metric as described by Moffat [8].

4. RESULTS

The disagreement between each pair of measures can be seen in Table 2. In some cases (such as AP vs RR), the disagreement is not surprising, but in other cases where one would expect metrics to agree, they do not–such as the high disagreement with SNDCG and most other metrics. Also of note is the high disagreement between RBP variants–although this is not surprising, it is worth mentioning that simply changing the discount parameter dramatically affects the behaviour of the metric.

4.1 Realistic disagreement

The disagreement percentages shown in Table 2 are assuming that each possible ranking with $1 \leq k \leq 10$ can be returned by some system. However, retrieval systems aren't random; the results returned depend on collection statistics and other features which are not independent of relevance. To investigate whether these vectors are reasonable, we produced the $k = 10$ bitstrings for all runs submitted to the TREC4-8 adhoc tracks, and the 2005 and 2006 TREC robust tracks. All of the 1024 possible vectors were returned by real runs submitted to the track. This validates this exhaustive method of investigation.

4.2 Properties of metrics

To investigate whether the properties of effectiveness measures affect the disagreement between metrics, each disagreement score was sorted in to one of three buckets for each property: *has-prop*, where both metrics have the property; *no-prop*, where neither metric has the property, and *cross-prop*, where one metric has the property, and one does not. Table 3 shows the mean agreement between metrics in each bucket. Surprisingly, a single property is not enough to determine whether two metrics agree. Of note is the case of metrics which are not bounded, but as Table 1 shows, there are only two measures in the *no-prop* category for that property. With the exception of monotonicity, boundedness, and realisability, the *has-prop* category contains the lowest mean disagreement. However, a one way ANOVA test with a Tukey post-hoc analysis shows only the convergent/has-prop group and the realisable/no-prop group to be statistically significant predictors of agreement.

4.3 Task behaviour

To investigate whether metrics disagreed in a predictable way that might be correlated with a task goal, all disagreeing pairs of ranked lists from each pair of metrics were examined to find out which metric preferred: *recall*, the list with the most relevant documents (ignoring ties); *first-rel*, the list with the first relevant document (ignoring ties); and *first-irrel*, the list with the first irrelevant document (ignoring ties). Some selected results are in Table 4. Each pairing is divided by the space in the table. When AP and DCG disagree about which ranked list is better, this table shows that in 52.41% of disagreeing cases, AP will prefer the list with more relevant documents, while DCG prefers the higher recall list only 22.10% of the time. These numbers do not sum to 100%, as sometimes there is a tie in the recall between two ranked lists, and sometimes a disagreement means that one of the two metrics gives the same score to two ranked lists (as is the case with AP when comparing $\{1,0,0,0\}$ and $\{0,1,0,1\}$ when $R = 10$).

This approach provides an alternative way to think about metric selection. When choosing say between AP and DCG (two metrics with strong agreement), this approach allows us to drill down into the cases where they disagree, and decide what type of task we favour most. If we prefer finding a relevant document quickest, then Table 4 suggests that DCG is the metric to choose, with 71.63% of cases preferring the ranking with the highest ranked initial relevant document. Alternatively, if recall is preferred, then AP is the better metric. Note that although Table 2 shows only a 1.32% disagreement between AP and DCG, this is still over 42,000 pairs of ranked lists for different values of k and R.

The difference between the user model of RBP and DCG can also be seen in this data–DCG prefers the list with the earliest relevant document more often than RBP. And, as the weighting factor in RBP decreases, the metric prefers earlier relevant documents. Conversely, as the weight in RBP increases, the list with higher recall is preferred.

5. CONCLUSION

In this work we have introduced a novel strategy for investigating the disagreement between effectiveness metrics–by counting and examining the pairs of hypothetical rankings where the metrics disagree with each other. We validated our strategy by demonstrating that all possible rankings of 10 binary relevant documents have appeared in search results submitted to two of the TREC tracks, and we performed an initial investigation into whether the properties of effectiveness measures can be used to predict agreement. We found that two of the properties appeared to be weak predictors of agreement between metrics. Then, we used a feature-based approach to investigate whether the disagreement between a pair of metrics could be described in terms of task features. This approach allows statements like "if I select AP over DCG, I am preferring recall over highly ranked documents" to be made, allowing fine-grained selection between metrics.

6. FUTURE WORK

As mentioned above, search engines do not produce a random result set. Although all possible rankings for $k = 10$ did appear in real search results during the TREC adhoc and robust tracks, the frequency with which each ranking appears is not uniform. It would be valuable to examine the likelihood of disagreement that each result list has. It is possible that the gain vectors produced by systems could be used to determine how contentious they are–how sensitive evaluation would be to metric selection.

An obvious extension to Section 4.2 is to consider multiple groupings of properties. Although it seems that individual properties of effectiveness measures are not enough to predict agreement, perhaps some combination of the properties might be. In future work, we intend to include more variants of metrics, such as alternative discounts and gain functions. This may lead to discovery of further properties.

In Section 4.3, we only consider three features of ranked lists that map to task goals, but there are many more we could consider (such as longest consecutive run of relevant documents, or lowest ranked relevant document). Additionally, user preference experiments could be constructed using pairs of vectors where metrics disagree.

As noted in Section 4, many of the DCG variants completely agree on individual ranked lists, as all that changes is the normalisation. However, if multiple queries (and therefore multiple ranked lists) are used for evaluation–as in the case of TREC tracks–then different normalisation strategies may well cause further disagreement between metrics.

Finally, this work has only considered the binary relevance case. Many of these metrics behave differently when there are multiple grades of relevance. An important next step is to repeat this analysis using graded relevance.

7. REFERENCES

[1] A. Al-Maskari, M. Sanderson, and P. Clough. The relationship between IR effectiveness measures and user satisfaction. In *Proc SIGIR*, pages 773–774, 2007.

[2] J. A. Aslam, E. Yilmaz, and V. Pavlu. The maximum entropy method for analyzing retrieval measures. In *Proc SIGIR*, pages 27–34, 2005.

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

[4] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proc CIKM*, pages 621–630, 2009.

[5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, Oct. 2002.

[6] E. Kanoulas and J. A. Aslam. Empirical justification of the gain and discount function for nDCG. In *Proc. CIKM*, pages 611–620, 2009.

[7] S. Mizzaro. The good, the bad, the difficult, and the easy: Something wrong with information retrieval evaluation? In *Proc. ECIR*, volume 4956, pages 642–646, 2008.

[8] A. Moffat. Seven numeric properties of effectiveness metrics. In *AIRS*, volume 8281, pages 1–12, 2013.

[9] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Syst.*, 27(1):2:1–2:27, Dec. 2008.

[10] P. Thomas, D. Hawking, and T. Jones. What deliberately degrading search quality tells us about discount functions. In *Proc SIGIR*, pages 1107–1108, 2011.