Retrieval Consistency in the Presence of Query Variations

Peter Bailey
Microsoft
Canberra, Australia

Alistair Moffat
The University of Melbourne
Melbourne, Australia

Falk Scholer
RMIT University
Melbourne, Australia

Paul Thomas
Microsoft
Canberra, Australia

ABSTRACT

A search engine that can return the ideal results for a person's information need, independent of the specific query that is used to express that need, would be preferable to one that is overly swayed by the individual terms used; search engines should be *consistent* in the presence of syntactic query variations responding to the same information need. In this paper we examine the retrieval consistency of a set of five systems responding to syntactic query variations over one hundred topics, working with the UQV100 test collection, and using Rank-Biased Overlap (RBO) relative to a centroid ranking over the query variations per topic as a measure of consistency. We also introduce a new data fusion algorithm, Rank-Biased Centroid (RBC), for constructing a centroid ranking over a set of rankings from query variations for a topic. RBC is compared with alternative data fusion algorithms.

Our results indicate that consistency is positively correlated to a moderate degree with "deep" relevance measures. However, it is only weakly correlated with "shallow" relevance measures, as well as measures of topic complexity and variety in query expression. These findings support the notion that consistency is an independent property of a search engine's retrieval effectiveness.

CCS CONCEPTS

•Information systems →Retrieval effectiveness; Test collections;

KEYWORDS

Test collections, retrieval consistency, semantic effectiveness

1 INTRODUCTION

Evaluating search effectiveness has several aspects. One aspect that has been especially popular is calculating average scores according to some relevance measure (such as NDCG or AP) over a set of topics and associated queries for some common corpus of information. In the batch evaluation methodology, different systems are compared using the same measure, and statistical tests are applied to determine whether the difference in performance is likely due to factors other than chance. Alternative methods for determining relevance include user studies, online interleaving, and A/B testing. What is important to note is that retrieval effectiveness encompasses more than just relevance.

Batch evaluation has typically used only a single query per topic, although a number of researchers working on early test collections advocated and explored the effect of using multiple queries per topic (see work by Spärck Jones and van Rijsbergen [35], Belkin et al. [5] and Buckley and Walz [7] among others). Recent work by Bailey et al. [3] and Koopman and Zuccon [22] has returned to this theme, resulting in new test collections with large numbers of query variations responding to each topic's information need.

The availability of such test collections allows us to consider a new dimension in assessing the retrieval effectiveness of search engines – namely, how *consistent* they are when returning results in response to query variations that address the same information need. The importance of consistency can be understood when we consider simple examples like mis-spellings. For example "*facebook*", "*facebok*", and "*faecbook*" pretty clearly all want to find the Facebook home page. Consistency also applies to more complex examples involving synonyms (for example, "*health benefits of vitamin c*" and "*health benefits of ascorbic acid*") or entirely rephrased needs (for example, "*how much does a raspberry pi cost*" and "*price of raspberry pi computer*"). In each of these cases, we can contemplate that there exists an ideal ranked set of relevant results drawn from the corpus. Test collections are premised on this principle, where the ideal set for a topic is discovered through judging a document pool formed from different rankings. An ideal search engine would return (only) this set of results given a query for a topic, and the difference in relevance from what is actually returned and this ideal ranking is captured by a relevance measure. Equally, given a set of syntactic query variations, an ideal search engine would return this ideal ranking of results, independent of the query variation.

Indeed, much research in information retrieval seeks to tackle exactly this problem of finding an ideal set of results without relying solely on the original query's syntactic expression. For instance, query re-writing techniques such as spelling correction [11], term stemming [24], query expansion [33], and query substitutions [19] are used to manipulate the user-entered query and thereby extract a better set of documents from the index. Stemming and stopword removal [23, 25] may also be used in indexing processes or within the matching algorithms at query execution time.

Two strands of research have investigated how to combine rankings to improve relevance effectiveness: data fusion (for example,

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

978-1-4503-5022-8/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3077136.3080839

Belkin et al. [6]), which merges rankings from different query representations; and distributed information retrieval or meta-search (for example, Callan [8]), which merges rankings from different underlying search engines or indexes. These techniques have been assessed principally from the standpoint of improving relevance overall, as measured by the relevance score of the resulting ranking.

People use many different expressions to describe the same information need (see, for example, Furnas et al. [15]). Even when refinding a single information resource, the same person may use different queries [37]. Bailey et al. [2] give evidence that the effect of query variation on relevance scores dwarfs that of system and topic effects. We believe that these findings make it important to consider new approaches to characterizing effectiveness, including ones that address query variation for a single information need.

We propose that the *consistency* in rankings of a system when faced with many different query variations for a single topic can be one such measure. The Rank-Biased Overlap measure developed by Webber et al. [38] is adopted to characterize consistency, and used with the UQV100 test collection [3] to investigate consistency across a set of five systems. Due to the scale of variations within UQV100 (19–101 unique query variations per topic, for 100 topics), each individual topic has a similar number of queries as might be found in an entire query-only processed test collection like the TREC 2014 Web track [10].

To determine relative consistency for a single system and topic combination, RBO requires us to declare some reference ranking against which the individual rankings for each query variation can be compared. We develop a new fusion algorithm, the Rank-Biased Centroid (RBC), drawing inspiration from both RBO and Rank-Biased Precision [27], to determine this reference ranking.

We consider these research questions in regard to RBC fusion:

RQ-F1 Do RBC rankings outperform the initial query rankings for a system?
RQ-F2 How does RBC compare to other data fusion algorithms in relevance effectiveness?
RQ-F3 Does combining both query variations and systems for RBC outperform query variations-only RBC?

Then, in connection with consistency, we ask:

RQ-C1 Do topics vary in consistency?
RQ-C2 Does consistency vary with the number of query variations or with changes in topic complexity?
RQ-C3 Do systems vary in consistency?
RQ-C4 Are increases in per-topic consistency for a system independent of increases in relevance for a system?

2 RELATED WORK

2.1 Data fusion

Data fusion – combining evidence from different sources – is a widely-studied problem. In IR fusion is typically applied when evidence from multiple ranked answer lists needs to be combined into a single ranked list, for example in meta-search, where results from multiple independent search systems are combined into a single ranking [1], and in multi-lingual retrieval, where results from searches across collections in different languages are combined into a single answer list [16].

Data fusion approaches can be broadly grouped into those that use the ranker's score (that is, the value assigned by a ranking function) of each document in a results list, and those that make use only of the rank position of each document in the answer list. Perhaps the most well-known approaches in the former category are by Fox and Shaw [13], including CombMAX, where the final score of a document is the maximum of the ranker scores that it received in any input ranked list; CombSUM, where the final document score is the total of the ranker's scores that it received in the input lists; and CombMNZ, where the final document score is calculated as in CombSUM but further multiplied by the number of input lists in which the document appears, thereby promoting those documents that were retrieved in multiple lists. The document ranker scores in the input ranked lists may also be normalized in different ways, including linear re-scaling into a chosen range [39], or by controlling an upper bound based on the sum or variance of the input scores [28].

For fusion based only on rank information, techniques from social choice theory such as the Borda count and Condorcet criterion have been applied. In the Borda count [1], each candidate (document) receives a score determined by how many other candidates were ranked lower, with these scores summed across all ballots (input lists). The Condorcet criterion instead determines an outcome based on which candidate achieves the highest number of wins based on pairwise comparisons with all other candidates [12, 29]

The impact of data fusion techniques on effectiveness can vary from case to case, leading Wu and McClean [40] to investigate approaches for predicting the performance impact of applying fusion. Their results showed that the selective application of fusion, based on features such as the number of component result lists and the overlap of items in these lists, can further enhance the positive impact on final retrieval effectiveness.

Prior work that has specifically considered data fusion in the context of multiple queries for the same underlying information need was carried out by Belkin et al. [5], who investigated the effects of combining five independent Boolean query formulations for ten TREC topics, and demonstrated that fusing results can substantially boost performance. Subsequent work using the TREC-2 collection further demonstrated that good methods for fusing the results of multiple queries can lead to results that are better than those of the best single query [6]. Pickens et al. [30] also confirmed that combining multiple queries for the same intent boosts effectiveness.

2.2 Measuring consistency

In IR, as in many other domains, it can be important to compare the similarity, or consistency, of groups of things. These groups may be conjoint (consisting of the same items) or disjoint (one group may include items that do not occur in the other group), and may be set-based (where there is no known or inferred ordering of the items) or ordered (where the sequence in which the items occur matters). Typical examples where one might wish to compare groups include measuring the similarity of the answer lists returned by two search engines in response to the same query; or the similarity of the effectiveness ranking of a set of several different retrieval systems, when evaluated over two different test collections. A wide range of list similarity measures have been proposed and applied.

One of the similarity measures most commonly used in IR is Kendall's τ, a rank correlation coefficient that calculates the normalized number of concordant pairs (items that are ordered the same in two rankings) minus discordant pairs (items that are ordered differently in the two rankings), resulting in a score between 1 (perfect agreement between the two rankings) and -1 (perfect disagreement) [9]. Kendall's τ is a measure that assumes conjoint ranked lists, that is, the lists are permutations of the same set of items; it is also an unweighted measure, where each pair contributes equally to the outcome, wherever it occurs in the ranking. However, when comparing ranked answer lists, items that are higher in the ranking are more important (users pay more attention to top-ranked results); similarly, when comparing rankings of system effectiveness scores, differences between the top systems are generally of greater interest than differences between systems that perform less well. TauAP [41] addresses a number of these weaknesses, while Rank-Biased Overlap (RBO) [38] addresses even more. RBO is a generalized measure of similarity between rankings based on a probabilistic user model, and readily handles non-conjoint lists. It applies a geometric sequence of weights to items in the lists; with the emphasis of the weighting adjustable via a user persistence parameter ϕ, which also determines the probability that the user will reach a certain rank. Inspired by RBO, Tan and Clarke [36] define a family of Maximized Effectiveness Difference (MED) measures, each based on an IR effectiveness metric (and hence a different underlying user model).

Jiang et al. [18] examine ranking consistency in web search from the basis of finding similar classes of queries (based on sharing common entity types in a knowledge base), and preserving the relative ordering of URL domains in the rankings for queries belonging to the same class, for example, people who are professional basketballers but also appear in movies. Large scale web log click data is used to derive their models of class similarity, based on URL patterns. Jiang et al. develop a consistency measure based on Kendall's τ across all pairs of queries belonging to the same class.

Finally, Zuccon et al. [43] present an evaluation framework using mean variance analysis over retrieval effectiveness, for both inter-topic and intra-topic sources of variation. Systems are preferred, all other things being equal, when one system is more stable than another in the presence of such variation.

3 RANK-BIASED CENTROIDS

As noted in the previous section, a range of methods have been proposed for constructing fused rankings, given an initial set of same-basis source rankings. In this section we introduce a further approach: the *rank-biased centroid*, or RBC.

3.1 User model for Borda fusion

To motivate the discussion, consider the four alternative rankings R1, R2, R3, and R4 shown in the left side of Figure 1, with each of the elements denoted by a letter of the alphabet. One run has ordered all of the seven different elements, while the other three are truncated and omit one or more of the items – a typical situation. Moreover, note that even if they are all of the same length, the runs might contain different subsets of a larger group of elements – they need not be permutations of each other. Finally, note that in the most general scenario, there are situations in which the provided rankings are prefixes of longer lists, themselves of unknown (and perhaps even infinite) length.

Rank	R1	R2	R3	R4	$\phi = 0.6$	$\phi = 0.8$	$\phi = 0.9$
1	A	B	A	G	A (0.89)	D (0.61)	D (0.35)
2	D	D	B	D	D (0.86)	A (0.50)	C (0.28)
3	B	E	D	E	B (0.78)	B (0.49)	A (0.27)
4	C	C	C	A	G (0.50)	C (0.37)	B (0.27)
5	G	–	G	F	E (0.31)	G (0.37)	G (0.23)
6	F	–	F	C	C (0.29)	E (0.31)	E (0.22)
7	–	–	E	–	F (0.11)	F (0.21)	F (0.18)

Figure 1: Example of RBC fusion: four example rankings (left); and three different fused orderings (right). Note that the RBC weights are shown to two decimal places only, and there are no score ties.

The Borda scoring process assigns a weight to item A of 7+7+4 = 18 (note that A does not appear in ranking R2), tying it with B, and placing it behind item D, which gets 6 + 6 + 6 + 5 = 23 points. The overall Borda ordering is D, A=B, C, G, E, F. In the Borda regime, swapping the two adjacent items at any pair of consecutive ranks gives one of the items a +1 score change, and the other item a -1 change. This occurs regardless of whether the swap takes place at rank 1, at rank 10, or at rank 100. That is, all binary item preferences as expressed in the visible input rankings are regarded as being of equal merit; and any preferences that may not have been surfaced (in the case that the provided rankings are prefixes) are ignored.

To create a user model that captures this behavior we can imagine a universe of agents, each of whom acts independently of the others, but follows the same simple rule: they randomly pick a depth d according to some probability distribution, they examine all of the input rankings to depth d (at most – but less if the rankings are shorter), and they sort the pool of items according to decreasing order of the number of times they saw each item in their set of length-d prefixes. The final fused ranking is then a probabilistic expectation over all agents of the orderings that were constructed.

Given this overall probabilistic structure, the Borda ordering is derived when the probability distribution used by the agents is taken to be $P(d = x) = 1/n$; that is, each agent is equally likely to select a prefix of any length between 1 and n, where n is the number of items. The Borda score for an item is then proportional to the expected value of the total number of times it was observed in the individual top-d sets of the probabilistic universe of agents.

3.2 An alternative weighting regime

Because there are many situations in which the supplied rankings are assumed to be prefixes of arbitrary-length ones, we contend that swaps near the heads of each of the rankings are somehow more indicative of preference than swaps deeper in the rankings. In the example shown in the left side of Figure 1, swapping A and D in ranking R1 has the same net effect on A's Borda score as does swapping A and F in ranking R4, but the latter swap might seem to be somewhat less damaging to A, since in R4 it has already been deprecated by the person or system that generated that ordering.

Instead of assigning a Borda weight of $(n - x + 1)/n$ (in a normalized sense) to each item at rank $1 \le x \le n$ when the rankings are over n items, we suggest that a geometrically decaying weight function be employed, with the distribution of d over depths x given not by $1/n$, but instead by $(1 - \phi)\phi^{x-1}$ for some value $0 \le \phi \le 1$ determined by considering the purpose for which the fused ranking is being constructed. The parameter ϕ is the *persistence*, or *patience* of the imagined universe of probabilistic agents; and use of a geometric sequence models the same behavior as is embedded in the effectiveness metric RBP [27] and in the rank correlation coefficient RBO [38] – namely, that the person examining the rankings always examines the first item in each, and thereafter proceeds from the i th to the $i + 1$ st with conditional probability ϕ, and ends their search at the i th with conditional probability $(1 - \phi)$. In an implementation the fused ranking is determined by assigning a weight of $(1 - \phi)\phi^{i-1}$ to each item at depth i in any of the rankings, and then summing over items and sorting by total weight.

There are a number of benefits of this proposed approach:

- as already motivated, greater emphasis is placed on the earlier preferences than on deeper ones in each ranking;
- an upper bound on the lengths of the rankings is not required, nor are the rankings required to be the same length (the Borda method shares this flexibility, albeit somewhat awkwardly);
- if further items are added at the tail of any of the rankings, the resultant item scores converge smoothly.

As extreme values, consider $\phi = 0$ and $\phi = 1$. When $\phi = 0$, the agents only ever examine the first item in each of the input rankings, and the fused output is by decreasing score of first preference; this is somewhat akin to a first-past-the-post election regime. When $\phi = 1$, each agent examines the whole of every list, and the fused ordering is determined by the number of lists that contain each item – a kind of "popularity count" of each item across the input sets. In between these extremes, the expected depth reached by the agents viewing the rankings is given by $1/(1 - \phi)$. For example, when $\phi = 0.9$, on average the first 10 items in each ranking are being used to contribute to the fused ordering; of course, in aggregate, across the whole universe of agents, all of the items in every ranking contribute to the overall outcome.

In practice, what this means is that different values of ϕ between 0 and 1 give rise to different fused orderings, balancing top-weightedness and exhaustivity. The right side of Figure 1 shows the orderings generated for the rankings R1, R2, R3, and R4, discussed earlier, for three different values of ϕ. The total weight associated with each item (to two decimals) is also shown. Note how item A is top-ranked when the ranking agents are relatively impatient, and (on average) abandon the ranking early ($\phi = 0.6$), but that if the fused ranking is assembled on a more patient basis ($\phi = 0.8$ and $\phi = 0.9$), items D and C become preferred, and A is demoted.

As with Borda fusion, unanimous preferences are respected: in the example, because C is below D in all of the four input rankings, it must also fall below D in the fused ranking, regardless of the value of ϕ. The differences that arise as ϕ varies are limited only to the elements where there is disagreement in the input rankings as to their respective ordering. These are, arguably, exactly the elements that we might be interested in focusing on.

3.3 Discussion

We have defined RBC in terms of a one-state user model [27]. Another way of looking at it is as an estimation of the normalized document scores used in CombMNZ and CombSUM. By assigning decreasingly small weights to documents further down the ranking, RBC can be viewed as seeking to approximate the long tail of document-query similarity scores generated by disjunctive ranked retrieval systems. Functions other than the geometric sequence might also be suitable for use, for example, Zipfian weightings.

4 FUSION OVER QUERY VARIATIONS

This section explores the practical benefit of fusing over query variations, and also shows that fusion over systems retains some of its power even after query variations have been incorporated.

4.1 The UQV100 collection

The UQV100 test collection is made up of 100 topics and associated information need statements, with approximately 100 individual query variations per topic; 10,835 in total [3]. When spelling correction and normalization are applied there are between 19 and 101 unique query variations per topic; 5,765 in total. There are also 55,587 relevance judgments available in regard to those 100 topics, covering ClueWeb12-CatB documents pooled from five systems, spanning three separate search engine code bases, and five different ranking algorithms [26]. We again employ the runs for those five contributing systems, anonymized here as Systems 1, 2, 3, 4, and 5. Due to some processing anomalies we observed in the run data, the overlapping set of unique query variations processed by all five systems contains 5,736 queries. Each system run contains a ranking of length 200 for each of those distinct queries. For definiteness we ordered the set of queries for each topic by decreasing frequency according to crowd-based process used to originally collect them [2], with ties broken randomly.

4.2 Fusion over query variations

Table 1 provides a detailed evaluation of approaches for fusion as applied to query variations. Each pane of the table gives results for one effectiveness metric, and within each pane the columns represent increasing numbers of query variations v (note that for $v \ge 20$, the legend $v = x$ indicates that as many as x query variations were used – some topics had fewer than the listed number of variations). Note that $v = 1$ makes use of the most frequently-suggested query for each of the UQV100 topics; $v = 2$ adds the second most frequently suggested one; and so on. Stepping across each row thus involves more and more input runs being used to form each output run, and as can be seen, effectiveness scores (with a few exceptions) increase. Many of the fusion methods give very similar effectiveness. Even so, there are some notable patterns:

- using as few as $v = 2$ query variations gives improved effectiveness (relative to the $v = 1$ baseline) for all metrics and all fusion methods;
- for all of the metrics, RBC with high values of p provides good fused outcomes, comparable with or better than those achieved by CombMNZ and Borda;
- for the two recall-based metrics, RBC-based fusion provides markedly better outcomes than Borda and CombMNZ;

Fusion	Number of variations per query					
	$v = 2$	$v = 4$	$v = 10$	$v = 20$	$v = 40$	$v = all$
RBC, $\phi = 0.9$	0.491	0.490	0.510	0.516	0.524	0.522
RBC, $\phi = 0.95$	0.497	0.504	0.516	0.526	0.529	0.526
RBC, $\phi = 0.98$	0.505	0.511	0.522	**0.535**	**0.533**	0.532
RBC, $\phi = 0.99$	**0.511**	0.520	0.525	**0.534**	**0.533**	**0.534**
Borda	0.511	**0.522**	**0.527**	0.534	0.532	**0.535**
CombMNZ	**0.513**	**0.521**	**0.527**	0.534	0.532	0.531

(a) RBP0.85, common baseline 0.474

Fusion	Number of variations per query					
	$v = 2$	$v = 4$	$v = 10$	$v = 20$	$v = 40$	$v = all$
RBC, $\phi = 0.9$	0.490	0.490	0.506	0.516	0.523	0.522
RBC, $\phi = 0.95$	0.496	0.504	0.514	0.526	0.528	0.527
RBC, $\phi = 0.98$	0.503	0.510	0.519	**0.533**	**0.533**	0.532
RBC, $\phi = 0.99$	**0.508**	**0.519**	0.523	**0.532**	**0.531**	**0.532**
Borda	0.507	**0.520**	**0.525**	0.530	0.528	**0.532**
CombMNZ	**0.509**	0.517	**0.524**	0.531	0.528	0.528

(b) INST, common baseline 0.471

Fusion	Number of variations per query					
	$v = 2$	$v = 4$	$v = 10$	$v = 20$	$v = 40$	$v = all$
RBC, $\phi = 0.9$	0.222	0.230	0.248	0.265	**0.288**†	**0.299**†
RBC, $\phi = 0.95$	0.223	0.234	**0.256**	**0.280**†	**0.297**†	**0.303**†
RBC, $\phi = 0.98$	0.225	0.239	**0.260**†	**0.275**†	0.281†	0.284†
RBC, $\phi = 0.99$	**0.226**†	**0.241**†	0.254†	0.264†	0.266†	0.270†
Borda	0.226	0.239	0.251	0.260	0.262	0.267
CombMNZ	**0.227**	**0.240**	0.252	0.260	0.273	0.266

(c) AP, common baseline 0.204

Fusion	Number of variations per query					
	$v = 2$	$v = 4$	$v = 10$	$v = 20$	$v = 40$	$v = all$
RBC, $\phi = 0.9$	0.437	0.454	0.484	0.505†	**0.539**†	**0.553**†
RBC, $\phi = 0.95$	0.438	0.458	**0.489**†	**0.519**†	**0.545**†	**0.554**†
RBC, $\phi = 0.98$	0.440	0.462	**0.492**†	**0.510**†	0.521†	0.525†
RBC, $\phi = 0.99$	**0.442**	**0.464**	0.481†	0.494†	0.499†	0.505†
Borda	0.442	**0.464**	0.478	0.489	0.493	0.502
CombMNZ	**0.442**	0.463	0.479	0.490	0.494	0.500

(d) NDCG, common baseline 0.409

Table 1: Fusion over query variations, average effectiveness across 100 UQV topics for runs generated from different numbers of query variations and according to different fusion approaches for System 1: (a) RBP0.85 scores; (b) INST scores; (c) AP scores; (d) NDCG scores. Query variations are sorted in decreasing order of occurrence frequency in the UQV100 collection. The baseline scores for $v = 1$ (that is, executing the single most popular query variation) are shown under each table. So that patterns of behavior can be seen, the two largest values in each column are highlighted in bold. Daggers indicate arrangements in which the RBC-based system was significantly better than the corresponding Borda run (one-sided paired t-tests with $p < 0.05$). No significant differences were detected for CombMNZ fusion, or using RBP0.85 or INST. Similar behavior was observed for other combinations of system and metric (not shown here).

Fusion	Metric			
	RBP0.85	INST	AP	NDCG
RBC, $\phi = 0.9$	0.503	0.501	0.217	0.441
RBC, $\phi = 0.95$	**0.508**	**0.506**	**0.219**†	0.442
RBC, $\phi = 0.98$	**0.506**	0.504	**0.220**†	**0.443**†
RBC, $\phi = 0.99$	0.505†	0.502	0.217†	0.440
Borda	0.503	0.500	0.215	0.440
CombMNZ	0.506	**0.505**	0.219†	**0.442**†

Table 2: Fusion over five different retrieval systems, based on one query variation ($v = 1$). All numbers are average effectiveness scores over the 100 topics in the UQV100 collection. Single-system scores for the four metrics are shown in the first two columns of Table 3. The largest two entries in each column are shown in bold. Daggers represent significance relative to Borda fusion ($p < 0.05$).

- moreover, the greater the number of query variations being fused, the smaller the value of ϕ needed to obtain those outcomes.

We also explored round-robin fusion and CombSUM fusion; the former was never competitive (and effectiveness decreased as query variations were added); and CombSUM typically gave performance slightly inferior to CombMNZ.

Metric	Initial, $v = 1$		Fused, $v = all$		Fused^2, $s = 5$	
	mean	max	mean	max	mean	gain
RBP0.85	0.474	0.487	0.530	0.557	0.559	+18%
INST	0.470	0.481	0.532	0.558	0.563	+20%
AP	0.190	0.204	0.268	0.303	0.303	+59%
NDCG	0.400	0.411	0.517	0.554	0.561	+41%

Table 3: Summary of effectiveness gains achieved by fusing first over query variations, and then second over systems. The first four data columns are mean and maximum average scores over five systems; the "gain" is relative to the initial system average in the first column. All fusion is carried out using RBC0.95.

4.3 Fusion over systems

Table 2 shows the outcome of applying fusion across the five systems used in our experiments. A single query is used in each input run ($v = 1$), and fusion applied to the five rankings for each topic. In this setting, all methods give comparable improvements in effectiveness, with Borda fusion marginally the worst of them.

4.4 Double fusion

Table 3 provides an overall summary of the effectiveness gains that can be achieved by fusing over query variations and then over systems, with RBC0.95 used at all fusing steps. Starting on the left, five systems each execute one query variation ($v = 1$) for each of the

Description	RBP0.85	INST	AP	NDCG
All queries ($v = all$)	0.405	0.394	0.151	0.336
First queries ($v = 1$)	0.474	0.471	0.204	0.409
Best query per topic	0.712	0.718	0.271	0.503

Table 4: Average metric scores for all queries per topic; for the most popular query per topic; and for per-metric per-topic "omniscient" query selections. System 1 is used throughout. These scores can be directly compared with those shown in the four panes of Table 1.

UQV100 topics; this can be regarded as being the starting baseline condition (no fusion performed) for both this table and Table 2. The five-way mean "average over 100 topics" and five-way maximum "average over 100 topics" values show typical behavior for five good retrieval systems when measured using a single query per topic. If each of those five systems is given more query variations, and generates a single fused run for each of the 100 topics as its output, the values in the middle pair of columns arise. Substantial performance improvements can be observed, and the means of the five "fusion over query variations" systems handsomely exceeds the best average score of the five original systems.

The third pair of columns in Table 3 then shows the outcome of fusing the five system runs generated after the query variations have been folded in. The mean scores shown (now with just a single ranking for each of the 100 topics) exceed the previous maximum scores in three of four cases, and exceed the middle-column mean scores in all four cases. The final column shows the end-to-end gain in effectiveness that has been achieved by the compound fusing ("initial mean" to "fused^2 mean"). That is, fusing first over query variations, and then over systems, gives rise to average effectiveness gains of 18% and higher. In terms of statistical significance, and looking at the various relativities summarized in Table 3:

- across the five systems and four metrics (twenty paired runs in total), the largest p-value computed by a two-tailed paired Student's t-test comparing the corresponding $v = 1$ and $v = all$ conditions was less than 0.005;
- when the five $v = all$ fused runs are compared with the final "fused^2" run, each metric yields one relatively large p-value, arising when the system that happens to be the "max" is compared with the final fused run ($p \approx 0.8$, 0.7, 0.9, and 0.3 respectively across the four metrics, with two different systems represented twice each as the "max" one), and a range of other smaller p-values, the largest of which was 0.059 (INST, comparing System 2 with $v = all$ against the final fused run), and the remainder of which were 0.01 or smaller.

That is, we are highly confident that fusion over queries helps retrieval effectiveness regardless of system and regardless of metric; and also confident that additional fusion across systems is also beneficial, helping ensure that the outcomes are as good as, or better than, what would have been attained if by chance we were already working with the best system for that metric.

4.5 An unrealistic target?

Hindsight is a wonderful guide, a fact that is also true in IR. Table 4 shows the result of a post-hoc evaluation of the runs generated for System 1. The first row shows the (unweighted) average metric scores across all of the UQV100 topics, and for each topic across the (approximately, on average) 55 distinct query variations. The second row shows the baseline effectiveness scores used in Table 1, arrived at by selecting the most popular of the query variations for each topic. The third row then shows results for four "oracle" query subsets, one for each metric, each incorporating (based post hoc on the relevance judgments and the computed metric scores) the "best" query variation for each of the UQV100 topics.

Comparing the first and second rows, the most frequently posed query generated by the crowd-workers for each topic obtains notably better effectiveness than the average of the variations. That difference is why we ordered the query variations as we did (Section 4.1). Comparing the second and third rows reveals a substantial further gap – for each of the topics and each of the metrics there are highly effective queries available within the sets created by the crowd-workers. For two of the metrics the single-query oracle runs are outperformed by the best of the fused approaches (Table 1), but for two metrics they are considerably better. Also worth noting is that the oracle runs have non-trivial differences, with different best queries arising for different metrics in many cases. Across the four metrics, a total of 193 best queries were identified.

5 CONSISTENCY DEFINED

5.1 Definition

As discussed in Section 2, Rank-Biased Overlap (RBO) [38] measures the top-weighted rank similarity between two non-conjoint indefinite rankings. As the rankings increase in similarity, especially towards the start of the rankings, the value of RBO trends towards 1.0. Due to the geometric sum of weights, governed by parameter ϕ, the total overlap score is bounded, ranging from 0.0 (no overlap) to 1.0 (total overlap).

To use RBO to measure consistency across query variations for a topic, we first select a common reference or objective ranking (the *centroid*) for each system-topic pair, making use of the RBC algorithm described in the previous sections. The persistence factor ϕ for RBC is set to 0.90, to mirror a user whose expected depth of examination into a ranked list is 10, a reasonably deep level of examination relative to standard web search; also, the UQV100 pooling ensured at least depth-10 judging for each query from each system. Different persistence factors could be selected, which would emphasize shallower or deeper probabilities of inspection of the lists forming the centroid or the depth of overlap. The centroid for all-systems might also have been considered; however we sought to measure a system's self-consistency, rather than with respect to a centroid that requires knowledge of other systems.

Given a centroid, we compute the RBO score for every query variation, using the same value of $\phi = 0.90$. Computation of RBO provides both a point-estimate, and a minimum, residual, and maximum value. Since system runs typically report 200 or more documents, when $\phi = 0.90$ the residual is less than 10^{-10}, and the point value is essentially equivalent to the minimum and maximum.

Formally, given a set of query variations $V_i = \{v_{i.1}, v_{i.2}, \ldots, v_{i.k}\}$ for a topic $t_i \in T = \{t_1, t_2, ..., t_n\}$, given a system S and its rankings for this set of variations $D_i = \{d_{v_{i.1}}, d_{v_{i.2}}, ..., d_{v_{i.k}}\}$, and given the corresponding RBC ranking for S and topic t_i, denoted $d_{i.\text{rbc}}$, we

measure *topic consistency* C_{t_i} with respect to S as

$$C_{t_i} = \frac{\sum_{q=1}^{k} \mathrm{RBO}(d_{v_{i.q}}, d_{i.\mathrm{rbc}})}{k}, \quad (1)$$

and the *collection consistency* C_T as the average topic consistency over the set of topics T, again with respect to S, as

$$C_T = \frac{\sum_{i=1}^{n} C_{t_i}}{n}. \quad (2)$$

That is, consistency is the average RBO score relative to the per-topic centroids generated for the system, expressed either as a set of per-topic scores, or aggregated over topics for a per-collection score, but always with regard to a system S. We can also speak of a *system's* consistency, which is simply topic (or collection) consistency for a particular system. Note that the near-zero RBO residual means that issues of averaging over RBO scores with different residuals can be ignored. We choose the average of averages for C_T because the number of variations per topic may vary, and topics with large numbers of unique variations should not unduly bias final scores.

Unlike for RBC, where we explored the consequence of adding more variations and thus needed an ordering, for consistency we use all unique query expressions without repeats (unweighted). In UQV100, each topic typically has a small number of commonly chosen variations which occur multiple times, and a large number of variations that occur only once. We wished to avoid biasing the consistency measure unduly by counting the contribution of the more popular variations multiple times. The choice of unweighted query variations also reinforces the decision to compute average of averages for C_T; in UQV100 the number of unique query variations per topic ranges from 19 to 101, so double averaging helps avoid undue influence from the topics with more diverse variations.

5.2 Why this definition of consistency?

Test collection-based evaluation reduces many sources of variance that occur in information seeking in the wild to a level that is tractable from the standpoint of statistical analysis. Our definition of consistency is predicated on having test collections that embody some plausible set of query variations per information need. We do not claim that it can address all possible sources of, or needs for, desirable consistency (or inconsistency) in information seeking.

Our definition of consistency might be brought into question by queries that exhibit extrinsic diversity [31] or intrinsic diversity [32] or searching as learning [14]. Extrinsic diversity occurs when a query has many different information needs that might be associated with it, while intrinsic diversity addresses cases where there are multiple sub-tasks associated in satisfying the information need. Searching as learning involves evolving query expression throughout a session. For cases involving extrinsic diversity, test collections without query variations typically declare one information need and judge relevance with respect to that information need; other plausible information needs are ignored. From the standpoint of assessing consistency, our approach is the same and has the same flaw of ignoring other information needs. For cases involving intrinsic diversity and searching as learning, approaches have been developed that target aspects of such complex evaluation situations, including TREC's Web track's Diversity task [10] and the Session [20] and Tasks [42] tracks. For assessing just consistency, we suggest that test collections should involve more narrow information needs, with an emphasis on developing specific unambiguous topic statements. Despite this, we believe that consistency is also important for systems that are able to accurately detect and respond to intrinsic diversity queries, and just as with narrower information needs, there will be a wide range of query variations for an information need that is intrinsically diverse.

One more question the reader might have is why measure consistency at all, and why not just go straight to average relevance. Two issues arise: first, relevance judgments are a substantially more expensive resource to accumulate, particularly when dealing with test collections with thousands of query variations. Second, although two rankings may have identical relevance scores, they may be completely different. Consider an information need such as "*You are worried about the prevalence of fake news, and decide to find authoritative newspapers to read instead, just like people did last century.*" Now consider two rankings in response to two query variations, one of which lists [theguardian.com, zeit.de, ft.com, washingtonpost.com] and one of which lists [wsj.com, lemonde.fr, nytimes.com, theglobeandmail.com]. From a relevance standpoint, these rankings are effectively identical, but from a consistency perspective, they share nothing. If we accept that a searcher cares about refinding the same information for an information need, even if they forget the precise query variation they used previously [37], then it is clear that being able to quantify consistency in rankings is not captured by relevance equivalence alone.

6 ANALYSIS OF CONSISTENCY

6.1 Consistency and topics

We address **RQ-C1** by assessing C_{t_i} over the five contributed runs described in Section 4.1. **RQ-C1** asks whether topics vary in consistency, and to do this we plot the average and standard deviation of C_{t_i} against all 100 topics of UQV100, sorted by increasing C_{t_i}. We characterize this in two ways: Figure 2 shows the results for System 1 (where standard deviation is of the RBO scores per variation for the topic), while Figure 3 shows the results when aggregating over the C_{t_i} scores obtained from the five systems. Even with the large standard deviations that can be observed in both plots (while being clearly smaller in the second), we can conclude that different topics have different consistency. For persistence $\phi = 0.9$, approximately 25% of topics have consistency under 0.25, while around 12% have consistency greater than 0.5. We also observe that some topics have great variation in their consistency scores, and others much less; and, overall, that consistency does indeed vary across topics.

6.2 Consistency and topic attributes

In the first part of **RQ-C2** we ask whether there is a relationship between the number of query variations per topic t_i and corresponding consistency scores C_{t_i}. Intuitively, we might expect that the more unique query variations there are for a topic, the lower the consistency scores. We address this aspect through a correlation analysis, using Spearman's ρ, a non-parametric rank correlation statistic. Unlike Pearson's product-moment coefficient statistic, this does not rely on the data having equal variance or having few to no outliers. A scatter plot, not shown, demonstrated that these

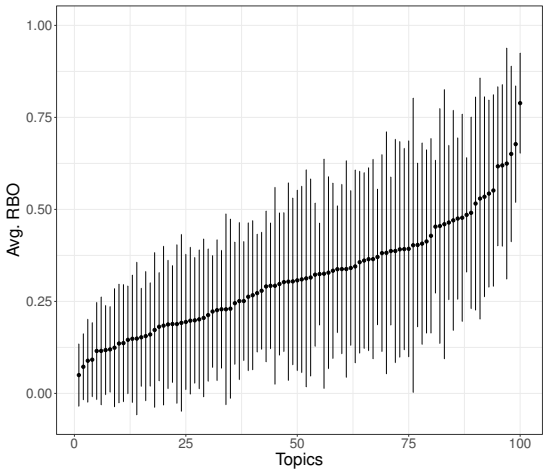

Figure 2: Consistency scores C_{t_i} for System 1, ordered by score, for 100 topics. Bars are ±1 s.d. of the underlying RBO values.

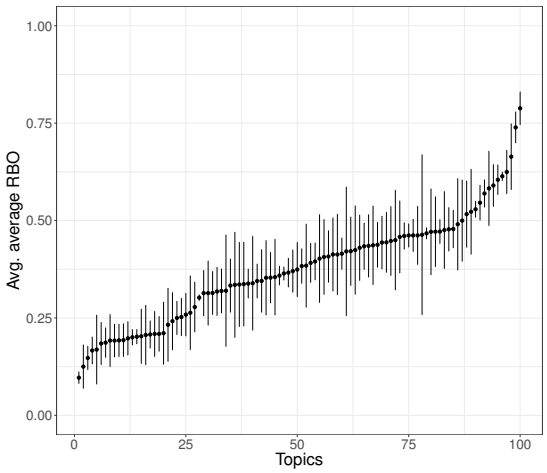

Figure 3: Average consistency scores C_{t_i} from five systems, ordered by increasing score, for 100 topics. Bars are ±1 s.d. Note that the topics may not be in the same order as in Figure 2.

requirements might not hold. In Table 5(a), we show the results for all systems. As expected, the direction of the association is negative (C_{t_i} goes down when the number of query variations goes up). However, the correlations are relatively weak weak magnitude ($0.3 < \rho < 0.5$) so although there is an association, it is not something we can reliably anticipate. The corresponding scatter plot of values is not shown for space reasons, but confirms that there is a broad range of consistency scores as the number of query variations per topic changes. These outcomes are a little surprising, suggesting that the causes of increased consistency are complex.

In the second part of **RQ-C2** we ask whether there is a relationship between the estimated topic complexity and corresponding consistency scores. Estimated topic complexity is available as the average of the estimates of the number of useful documents (and the number of queries) expected by each person providing a query

System	(a) Num. variations	(b) Est. docs	(c) Est. queries
1	−0.35‡	−0.30‡	−0.43‡
2	−0.43‡	−0.27‡	−0.43‡
3	−0.42‡	−0.35‡	−0.45‡
4	−0.36‡	−0.25‡	−0.39‡
5	−0.37‡	−0.36‡	−0.43‡

Table 5: Correlation measured using Spearman's ρ between C_{t_i} and: (a) number of unique query variations per topic; (b) average estimated useful documents per topic; and (c) average estimated useful queries per topic, for each (system, topic) pair. In all cases, $p < 0.01$(‡).

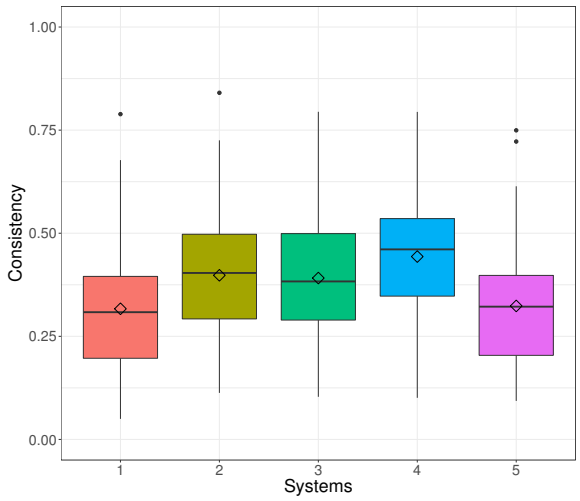

Figure 4: Consistency scores C_{t_i} for five systems and 100 topics. The diamond marks the median, and the horizontal line marks C_T.

variation for a topic description in UQV100. More complex topics have higher values for these two averaged estimates. Again, intuitively we might expect that the more complex a topic is, the lower the consistency score. As above, we assess the relationship using Spearman's ρ, for both the average estimated documents and average estimated queries per topic t_i and corresponding consistency scores C_{t_i}. Results are shown in Table 5(b) and (c), and just as before indicate a negative association, as surmised. However, once again the correlations are weak at best, at best of weak magnitude, meaning that increases in estimated topic complexity are only loosely associated with decreases in consistency. Interestingly, the estimates of the required number of queries all have a stronger correlation than the corresponding estimates of the required number of useful documents. This outcome might in part arise because the complexity estimate providers may be better at estimating changes in small numbers than in larger ones, but there are other possible explanations.

System	2	3	4	5
1	−9.756‡	−10.497‡	−14.362‡	−1.030
2		−0.930	−6.821‡	9.107‡
3			−8.035‡	8.490‡
4				12.330‡

Table 6: System differences measured using a paired Student's t-test between all pairs of systems over the corresponding C_{t_i} per topic. Values reported are the t statistic, $df = 99$ in all cases, and significant differences at $p < 0.05$ (†) and $p < 0.01$ (‡).

6.3 Consistency and systems

If consistency was identical across different retrieval systems, then it would be uninteresting when selecting an effective system. In **RQ-C3**, we examine the relationship between collection consistency C_T and the five systems which contributed to UQV100. In Figure 4 we report on the consistency scores for each system, using boxplots to show the spread of values. From this we can observe that Systems 1 and 5 are very similar to each other ($C_T \simeq 0.32$); Systems 2 and 3 are very similar to each other and more consistent ($C_T \simeq 0.40$), and System 4 is different and yet more consistent ($C_T = 0.44$). Using a two-tailed paired Student's t-test, we also assessed each pair of systems; Table 6 confirms our observations.

Another way of understanding the relationship between consistency and systems is to treat each topic as an "assessor" and each system as the "subject" being assessed with regards to its degree of consistency. Since topics in UQV100 contain a similar number of query variations (average of 55 per topic) as many existing test collections have queries, and past practice has been to examine rank order correlation of systems by comparing sets of systems across two or more collections, we will adopt a similar method here. We assess how similar the relative ordering of systems is using one-way Intraclass Correlation [4], Kendall's Coefficient of Concordance [21] (commonly written as Kendall's W), and Krippendorff's α [17]. All of these measures address inter-assessor agreement, for three or more assessors, and can accommodate ordinal data (and for ICC and α, interval or ratio data). In Table 7, we report the results for these measures of rank agreement, where the score is the topic consistency C_{t_i}. We use these measures rather than pair-wise comparisons of topics using Kendall's τ, since we have 100 topic "assessors" involved, and the measures allow us to consider multiple "assessors" with a single test statistic, while Kendall's τ only supports two "assessors". Due to the differences between ICC and Kendall's W, it is expected that ICC scores may be lower than Kendall's W scores over the same data, since it considers not just relative rank order but also the magnitude of differences, as discussed by Sheskin [34]. In all three measures, 1 indicates perfect agreement among the assessors, and 0 indicates no agreement beyond what would be expected by chance. Both ICC and α can report small negative values, which also signify no agreement. The values from ICC and Krippendorff's α both indicate there is a very low degree of inter-assessor agreement in rank ordering the systems by consistency; and although Kendall's W is 0.491 for C_{t_i}, this is still a relatively low degree of agreement.

From these two analyses, we conclude that although these systems do have different overall collection consistency C_T, they are not systematically ordered on the basis of topic consistency C_{t_i}. That is, for one topic a particular system may have high consistency, while for another topic a completely different system may have high consistency, and the earlier system may have low consistency.

Agreement	(a) ICC	(b) Kendall's W	(c) Krippendorff's α
C_{t_i}	0.111‡	0.491‡	0.090
NDCG	−0.002	0.169‡	−0.003
INST	−0.005	0.033†	−0.006

Table 7: Rank agreement over all five systems measured using (a) Intraclass Correlation; (b) Kendall's W; and (c) Krippendorff's α. For ICC and W, significance is reported as $p < 0.05$(†), and $p < 0.01$(‡); for α it is not reported. The top row uses C_{t_i} as the ranking score for each system; the bottom two rows use NDCG and INST (averaged by topic, as for C_{t_i}).

System	AP	NDCG	Q	RBP	INST
1	0.61‡	0.68‡	0.56‡	0.32‡	0.35‡
2	0.62‡	0.70‡	0.58‡	0.33‡	0.37‡
3	0.55‡	0.62‡	0.53‡	0.32‡	0.36‡
4	0.59‡	0.65‡	0.55‡	0.39‡	0.40‡
5	0.59‡	0.65‡	0.53‡	0.25†	0.30‡

Table 8: Correlation measured using Spearman's ρ between topic consistency C_{t_i} and corresponding relevance measures (AP, NDCG, Q measure, RBP0.85, and INST), for each (system, topic) pair. In all cases, there is a significant correlation, with $p < 0.01$ (‡), except for System 5 and RBP, significant only at $p < 0.05$ (†).

6.4 Consistency and relevance

Our last investigation, addressing **RQ-C4**, concerns the relationship between consistency and relevance. The construction of UQV100 guaranteed relevance judgments to at least depth 10 for all query variations for the five systems being analyzed. Thus we are able to explore the degree of correlation between consistency and relevance, across a range of relevance measures, including AP, NDCG, Q measure, RBP0.85, and INST. The results, calculated using Spearman's ρ, are displayed in Table 8. While there is moderate correlation for the "deep" relevance measures (AP, NDCG, and Q), there is only weak correlation for the "shallow" relevance measures (RBP0.85 and INST). Scatter plots of the data, not shown, indicate considerable spread of scores for all measures as C_{t_i} increases.

As a comparison with consistency, we repeat the topics-as-"assessors" inter-assessor agreement analysis, with results reported for average-by-topic NDCG and INST scores in rows two and three of Table 7. While any degree of agreement was only just observable for consistency, with these relevance measures, any agreement on the ordering of systems by relevance is effectively random.

7 CONCLUSIONS

Consistency – the ability to give similar results for a topic even when presented with different queries – is desirable for search engines in a variety of circumstances. We have defined a consistency

measure and explored it across a set of 5,736 query variations across 100 topics, using a novel relevance-based centroid algorithm.

The RBC algorithm for fusing rankings has the benefit of incorporating a persistence parameter that allows modeling of different depths of attention into rankings, and adds another strand to the "RB-" family. RBC is competitive or better than existing algorithms, and like Borda count has no reliance on system scores. We confirmed previous findings that data fusion over queries and over systems is beneficial, and fusion over both is even better. With the oracle runs we have also demonstrated that substantially better effectiveness performance is possible, at least hypothetically.

Based on the various analyses, we can also state that the consistency measure informs us about something different to existing measures. Consistency varies by topic and by system, tends to decrease as topic complexity and the number of query variations increases, and has weak-to-moderate correlations with several relevance measures. However, in no circumstance is consistency strongly correlated with any of these existing test collection dimensions, confirming that it measures a different property altogether. Neither the measures of consistency nor relevance reliably order the five systems over the UQV100 topics, indicating there are no clear system winners or losers for this test collection on these dimensions of effectiveness when examined topic by topic.

More investigation is required into the nature of consistency and its effect on perceptions of the retrieval effectiveness of search systems. Such work might include user studies, low-level analysis of the root causes of variable ranking within one or more systems, and broadening the current analysis to similar test collections with multiple query variations per topic.

Acknowledgment This work was supported by the Australian Research Council's *Discovery Projects* Scheme (projects DP110101934 and DP140102655). Matt Crane, Xiaolu Lu, David Maxwell, and Andrew Trotman assisted greatly, providing the system runs that were analyzed. The UQV100 judgments were generated using resources provided by Microsoft.

REFERENCES

[1] J. A. Aslam and M. Montague. 2001. Models for metasearch. In *Proc. SIGIR*. ACM, 276–284.

[2] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2015. User variability and IR system evaluation. In *Proc. SIGIR*. 625–634.

[3] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A test collection with query variability. In *Proc. SIGIR*. 725–728. Public data: http://dx.doi.org/10.4225/49/5726E597B8376.

[4] J. J. Bartko. 1966. The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19, 1 (1966), 3–11.

[5] N. J. Belkin, C. Cool, W. B. Croft, and J. P. Callan. 1993. The effect of multiple query representations on information retrieval system performance. In *Proc. SIGIR*. 339–346.

[6] N. J. Belkin, P. Kantor, E. A. Fox, and J. A. Shaw. 1995. Combining the evidence of multiple query representations for information retrieval. *Inf. Proc. & Man.* 31, 3 (1995), 431–448.

[7] C. Buckley and J. Walz. 1999. The TREC-8 query track. In *Proc. TREC*.

[8] J. Callan. 2002. Distributed information retrieval. In *Advances in Information Retrieval*. Springer, 127–150.

[9] B. Carterette. 2009. On rank correlation and the distance between rankings. In *Proc. SIGIR*. 436–443.

[10] K. Collins-Thompson, C. Macdonald, P. N. Bennett, F. Diaz, and E. M. Voorhees. 2014. TREC 2014 web track overview. In *Proc. TREC*.

[11] S. Cucerzan and E. Brill. 2004. Spelling correction as an iterative process that exploits the collective knowledge of web users. In *Proc. EMNLP*. 293–300.

[12] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. 2001. Rank aggregation methods for the web. In *Proc. WWW*. 613–622.

[13] E. A. Fox and J. Shaw. 1993. Combination of multiple searches. In *Proc. TREC*. 243–252.

[14] L. Freund, H. O'Brien, and R. Kopak. 2014. Getting the big picture: Supporting comprehension and learning in search. In *Proc. Searching As Learning (SAL) Workshop*.

[15] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. 1987. The vocabulary problem in human-system communication. *Comm. ACM* 30, 11 (1987), 964–971.

[16] F. C. Gey, N. Kando, and C. Peters. 2005. Cross-language information retrieval: The way ahead. *Inf. Proc. & Man.* 41, 3 (2005), 415–431.

[17] A. F. Hayes and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Commun. Methods and Measures* 1, 1 (2007), 77–89.

[18] J.-Y. Jiang, J. Liu, C.-Y. Lin, and P.-J. Cheng. 2015. Improving ranking consistency for web search by leveraging a knowledge base and search logs. In *Proc. CIKM*. 1441–1450.

[19] R. Jones, B. Rey, O. Madani, and W. Greiner. 2006. Generating query substitutions. In *Proc. WWW*. 387–396.

[20] E. Kanoulas, B. Carterette, M. Hall, P. Clough, and M. Sanderson. 2011. Overview of the TREC 2011 session track. In *Proc. TREC*.

[21] M. G. Kendall and B. B. Smith. 1939. The problem of *m* rankings. *Annals of Mathematical Statistics* 10, 3 (1939), 275–287.

[22] B. Koopman and G. Zuccon. 2016. A test collection for matching patients to clinical trials. In *Proc. SIGIR*. 669–672.

[23] R. T.-W. Lo, B. He, and I. Ounis. 2005. Automatically building a stopword list for an information retrieval system. *J. Dig. Inf. Man.* 3, 1 (2005), 3–8.

[24] J. B. Lovins. 1968. *Development of a stemming algorithm.* MIT Information Processing Group, Electronic Systems Laboratory Cambridge.

[25] H. P. Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.* 1, 4 (1957), 309–317.

[26] A. Moffat. 2016. Judgment pool effects caused by query variations. In *Proc. Aust. Doc. Comp. Symp.* 65–68.

[27] A. Moffat and J. Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. Inf. Sys.* 27, 1 (2008), 2.1–2.27.

[28] M. Montague and J. A. Aslam. 2001. Relevance score normalization for metasearch. In *Proc. CIKM*. 427–433.

[29] M. Montague and J. A. Aslam. 2002. Condorcet fusion for improved retrieval. In *Proc. CIKM*. 538–548.

[30] J. Pickens, G. Golovchinsky, C. Shah, P. Qvarfordt, and M. Back. 2008. Algorithmic mediation for collaborative exploratory search. In *Proc. SIGIR*. 315–322.

[31] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. 2009. Redundancy, diversity and interdependent document relevance. *SIGIR Forum* 43, 2 (2009), 46–52.

[32] K. Raman, P. N. Bennett, and K. Collins-Thompson. 2013. Toward whole-session relevance: Exploring intrinsic diversity in web search. In *Proc. SIGIR*. 463–472.

[33] S. E. Robertson. 1990. On term selection for query expansion. *J. Documentation* 46, 4 (1990), 359–364.

[34] D. J. Sheskin. 2003. *Handbook of Parametric and Nonparametric Statistical Procedures.* CRC Press.

[35] K. Spärck Jones and C. J. van Rijsbergen. 1975. *Report on the need for and the provision of an "ideal" information retrieval test collection.* Technical Report 5266. Computer Laboratory, University of Cambridge. British Library Research and Development Report.

[36] L. Tan and C. L. A. Clarke. 2015. A family of rank similarity measures based on maximized effectiveness difference. *IEEE Trans. Know. Data Eng.* 27, 11 (2015), 2865–2877.

[37] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. 2007. Information re-retrieval: Repeat queries in Yahoo's logs. In *Proc. SIGIR*. 151–158.

[38] W. Webber, A. Moffat, and J. Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Sys.* 28, 4 (2010), 20.1–20.38.

[39] M. Wu, D. Hawking, A. Turpin, and F. Scholer. 2012. Using anchor text for homepage and topic distillation search tasks. *JASIST* 63, 6 (2012), 1235–1255.

[40] S. Wu and S. McClean. 2006. Performance prediction of data fusion for information retrieval. *Inf. Proc. & Man.* 42 (2006), 899–915.

[41] E. Yilmaz, J. A. Aslam, and S. Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proc. SIGIR*. 587–594.

[42] E. Yilmaz, M. Verma, R. Mehrotra, E. Kanoulas, B. Carterette, and N. Craswell. 2015. Overview of the TREC 2015 tasks track. In *Proc. TREC*.

[43] G. Zuccon, J. Palotti, and A. Hanbury. 2016. Query variations and their effect on comparing information retrieval systems. In *Proc. CIKM*. 691–700.