Reusable Test Collections Through Experimental Design

Ben Carterette*, Evangelos Kanoulas†, Virgil Pavlu‡, Hui Fang⋆
carteret@cis.udel.edu, e.kanoulas@sheffield.ac.uk, vip@ccs.neu.edu, hfang@ece.udel.edu

* Department of Computer & Information Sciences, University of Delaware, Newark, DE
† Information Studies Department, University of Sheffield, Sheffield, UK
‡ College of Computer and Information Science, Northeastern University, Boston, MA
⋆ Department of Computer & Electrical Engineering, University of Delaware, Newark, DE

ABSTRACT

Portable, reusable test collections are a vital part of research and development in information retrieval. Reusability is difficult to assess, however. The standard approach—simulating judgment collection when groups of systems are held out, then evaluating those held-out systems—only works when there is a large set of relevance judgments to draw on during the simulation. As test collections adapt to larger and larger corpora, it becomes less and less likely that there will be sufficient judgments for such simulation experiments. Thus we propose a methodology for information retrieval experimentation that collects evidence for or against the reusability of a test collection *while* judgments are being made. Using this methodology along with the appropriate statistical analyses, researchers will be able to estimate the reusability of their test collections while building them and implement "course corrections" if the collection does not seem to be achieving desired levels of reusability. We show the robustness of our design to inherent sources of variance, and provide a description of an actual implementation of the framework for creating a large test collection.

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval] Performance Evaluation

General Terms: Experimentation, Measurement

Keywords: information retrieval, test collections, reusability, evaluation

1. INTRODUCTION

Test collections are a vital part of research and development in information retrieval. They enable rapid development of new approaches to retrieval. They allow us to identify subtle distinctions between retrieval methods that could not be identified by users but that can add up to improved user experience over time. They support feature selection and parameter tuning by allowing us to efficiently test many possible combinations and values.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

Unfortunately, test collections are expensive. They require judgments of the relevance of individual documents to topics in a sample. To properly control for variance, a test collection must have many topics and many judgments, and these require a great deal human effort. This expense makes *reusability* desirable: the cost of a test collection can be justified by the fact that it is amortized over many uses.

Constructing reusable test collections is difficult. The relevance judgments must be complete enough that future users of that collection can have confidence that their systems will be accurately evaluated. The majority of reusable test collections in the field exist as a result of the efforts of the organizers and participants of TREC (the Text REtrieval Conference), CLEF (the Cross Language Evaluation Forum), NTCIR (NII Test Collections for IR), and INEX (INitiative for the Evaluation of XML retrieval). These test collections arose by conducting an experiment to evaluate different approaches to a particular retrieval problem, and their reusability is a function of their large size and the diversity of approaches that were included in the experiment.

The standard experimental design for IR evaluation is a simple repeated-measures design, in which experimental units are topics/queries, treatments are systems, and each system provides ranked results for each query. This is the design that has been used for virtually every TREC, CLEF, NTCIR, and INEX track that has resulted in the release of a test collection. Measurements on experimental units are evaluation measures such as average precision (AP) calculated over relevance judgments; judging a pool of documents retrieved by the participating systems ensures that the measurements will be as accurate as possible.

Reusability emerges as a result of using a large and diverse set of retrieval systems and making sure they are judged to a substantial depth using pooling: there are simply so many judgments that it is unlikely any new system will ever be developed that does not retrieve many of the same documents that were judged as part of the original experiment. But as test collections grow larger and larger, pooling becomes more infeasible. Furthermore, recent work suggests that, for an experiment like that described above, it is actually more cost-effective to use many queries with very few judgments each. Thus TREC has begun adopting alternatives to pooling [15]: statistical sampling, which attempts to pick out the judgments that will result in a low-variance unbiased estimator [2], or algorithmic approaches that try to pick out the judgments that will reduce variance regardless of bias introduced [6]. But while these are more cost-effective for answering the original evaluation question, it is not at all

clear that they are cost-effective in the sense of producing test collections that can be reused many times. Because there are many fewer judgments, it is much more likely that a future system will retrieve many documents that were not judged, and therefore much more likely that we will not be able to accurately measure the performance of that system.

Our goal is to elevate reusability to a basic consideration along with evaluation. We do that by proposing an experimental design that collects evidence for or against reusability *while* judgments are being collected. The method that is used to select judgments does not matter, but the design is tied to the notion that more queries with fewer judgments is the correct way to build a test collection. It relies on having a large number of queries that can be partitioned into a combinatorial number of blocks.

In Section 2 we define what it means for a test collection to be reusable and discuss previous work on the topic. In Section 3, the main body of this work, we describe our design and the statistical analyses that it supports, and anticipate and answer some questions about its validity. In Section 5 we demonstrate the use of the design and analysis in the construction of an actual large test collection.

2. TEST COLLECTION REUSABILITY

A test collection consists of a corpus of documents, a set of topics that are representative of a particular task, and judgments of relevance of documents to topics. These judgments are generally taken from a set of retrieval systems performing the task. We define reusability as follows: A test collection is reusable *if and only if* we can use it for precise measurements of the performance of systems that did not contribute to its judgments. By "precise" we mean that the measurements fall within some given error intervals with high probability. By "systems that did not contribute judgments" we mean systems that are likely to be developed with current technology—it is always possible that new, unforeseen technology could produce retrieval systems that are both good and unlike anything seen before; since we will never be able to predict such cases, we do not want to tie reusability to them too much.

Test collections are used for many purposes beyond simple evaluation. Furthermore, evaluation comes in many different flavors. Below we discuss some aspects of reusability and previous work on this topic.

2.1 Applications of Test Collections

In addition to evaluation, test collections are used for training and optimization, including model selection, feature selection, and parameter tuning, for failure analysis, for data exploration, and many other purposes. While our focus is on evaluation, these other uses are important. Some can be seen as being related to evaluation: optimization uses an objective function based on an evaluation measure; the goal of failure analysis is to find reasons for an evaluation measure being different than expected.

2.2 Evaluating Reusability

The question of reusability has been studied primarily in the context of depth pooling. TREC and other fora form pools from the top documents retrieved by each submitted run for each topic; under the assumption that documents not highly ranked could be considered nonrelevant, test collections based on such pools are likely very reusable.

Harman [11] tested this by examining a pool formed by the documents in ranks 101-200 over the TREC-2 and TREC-3 collections. Her study showed up to 21% more relevant documents could be found. Along the same line, Zobel [17] extrapolated the number of relevant documents found by depth to suggest that there could be up to twice as many relevant documents in the collection as there are in the pool. To examine the effect of the missing relevant documents on new systems that had not contributed any documents to the formation of the pool, he performed a leave-one-*run*-out simulation. For each participating run, he removed all the documents it uniquely retrieved from the judgments and compared the evaluation over this reduced set of judgments to the evaluation with the full set. His study showed that the effect of the missing documents was minimal.

Voorhees adapted the leave-one-out methodology to leave out *all* of the runs contributed by a particular *site* at a time, under the assumption that runs submitted by the same participating site are similar enough that they retrieve very similar documents [14]. This leave-sites-out simulation has since become the standard approach to evaluating reusability.

Büttcher et al. [3] employed the leave-one-site-out over the TREC 2006 Terabyte collection and confirmed Zobel's conclusion. Further, the reusability of the collection by leaving out all manual runs was also tested. Given that manual runs are usually among the best performing ones, this did lead to somewhat different evaluation results.

Sakai [13] employed leave-one-site-out, take-just-one-site, and take-just-three-sites over TREC and NTCIR data. His goal was to identify the effects of missing judgments on a number of different evaluation metrics. He considered all pairs of runs over the full judgment set and found the number with statistically significant differences, then repeated the process with judgments obtained by one of the three aforementioned methods and counted the errors. The results demonstrate that while the rankings of systems over the full and reduced set of judgments are similar, missing relevant documents leads to many errors of commission, i.e. finding differences significant even though they are not.

Carterette et al. proposed that reusability should be evaluated in terms of the ability of the test collection to produce high confidence in evaluation results, specifically pairwise comparisons between systems [4] or width of confidence interval on an evaluation measure [7]. The former work used judgments from two systems to evaluate a larger set of 10 systems; the latter employed the leave-sites-out methodology discussed above to predict confidence interval width when evaluating new systems.

The simulation approaches above depend on having a fairly large number of judgments in the first place: any document that is selected for judging in the simulation phase must already have an actual judgment made by a human assessor. Without a fairly complete set of judgments it is likely that documents selected for judging will not actually have judgments; it is not possible to apply simulation to evaluate the reusability of TREC Million Query collections, for instance, because holding systems out would result in different documents being selected for judging than were originally judged for the track.

2.3 Types of Reusability

Based on the work above, we identify three types of analysis that test collections are used for in evaluation:

1. "within-site" analysis, in which a research/development site is conducting an experiment to determine which of several possible (internally-developed) systems to publish or deploy. We believe this is the most common use of test collections.
2. "between-site" analysis, in which one research/development site compares their results to those of another site, possibly relying on published results.
3. "participant comparison" analysis, in which a research/-development site compares their results to those of the systems that are on record as participating in a particular track or task.

"Site" is TREC terminology, but it can be defined loosely; within a particular setting, any group of systems that are similar in some sense could be considered a "site". We use the term in that general sense throughout this work.

Our goal is to develop a methodology that can be used to test all three types of reusability when simulation is impossible due to the process used to select documents to judge.

3. EXPERIMENTAL DESIGN

As discussed above, the standard design used in system-based IR evaluations is the repeated-measures design. This is appropriate for drawing conclusions about differences between systems, but it does not tell us anything about reusability. Furthermore, as discussed in Section 2.2, post-hoc evaluations of reusability are impossible when refinements to the implementation of the repeated-measures design such as statistical sampling or algorithmic selection were used.

In our design, *each* system is held out from *actual* judgment collection for some queries. After the judging is complete, "new" systems are constructed by putting together all the queries from which a system was held out and evaluating it with the judgments contributed by the non-held-out systems for those same queries. Note that this means the reusability experiment can be performed only once.

Our design is meant to serve two ends: to draw conclusions about differences between systems, and to draw conclusions about the future reusability of the test collection that will result. It is meant to be "fair" in the sense that each system contributes judgments to the same number of queries. Since it can introduce bias or variance depending on which systems are held out from which queries, it attempts to minimize/control that as much as possible by ensuring that no two systems are held out of the same queries consistently. The complete description follows.

3.1 Description of Design

We partition N topics into $b+1$ sets $T_0, T_1, \ldots, T_b$. The first set, T_0, consists of n topics to which all systems contribute judgments. This is the standard repeated-measures design to ensure that we can answer questions about differences between these systems. It provides a baseline for answering questions about reusability.

In each subsequent set, a subset of systems are held out during judgment collection for each topic. The held-out set is different for each topic. Choosing which systems to held out can be done by site (if multiple sites have contributed systems): if there are m sites, k are held out from each query in the set; which k to hold out can be determined using round robin. The total number of queries must be a multiple of $\binom{m}{k}$ to ensure that each site is held out of the same number of queries.

subset	topic	S_1	S_2	S_3	S_4	S_5	S_6
T_0	t_1	+	+	+	+	+	+
all-site	$\cdots$						
baseline	t_n	+	+	+	+	+	+
T_1	t_{n+1}	+	+	+	+	−	−
	t_{n+2}	+	+	+	−	+	−
	t_{n+3}	+	+	−	+	+	−
	t_{n+4}	+	−	+	+	+	−
	t_{n+5}	−	+	+	+	+	−
	t_{n+6}	+	+	+	−	−	+
	t_{n+7}	+	+	−	+	−	+
	t_{n+8}	+	−	+	+	−	+
	t_{n+9}	−	+	+	+	−	+
	t_{n+10}	+	+	−	−	+	+
	t_{n+11}	+	−	+	−	+	+
	t_{n+12}	−	+	+	−	+	+
	t_{n+13}	+	−	−	+	+	+
	t_{n+14}	−	+	−	+	+	+
	t_{n+15}	−	−	+	+	+	+
T_2	t_{n+16}	+	+	+	+	−	−
	$\cdots$	$\cdots$					
	t_{n+30}	−	−	+	+	+	+
T_3	$\cdots$						

Table 1: Illustration of proposed experimental design at the site level with $m = 6$ sites and $k = 2$ held out from each topic. Each column shows which topics a site contributed to. A + indicates that all of the sites' runs contributed judgments to the topic; − indicates that the sites' runs did not contribute judgments. Each subset $T_1 \ldots T_b$ has the same contribution pattern as subset T_1.

This design is essentially a standard randomized, repeated-measures block design in which blocks are defined by which sites have been held out; there are $\binom{m}{k}$ blocks and b observations in each block. Statistical tools such as mixed-effects ANOVA can be applied directly to answer questions about differences between individual systems. Answering questions about reusability will require some additional tools that we describe in the next section.

The design is illustrated in Table 1 to give a sense of how it provides data for each of our three types of reusability:

1. "within-site": Within each subset T_i, each site contributes to $\binom{m-1}{k}$ topics and is held out from $\binom{m-1}{k-1}$ topics. Thus in addition to the n topics that all sites contribute to, each site contributes to $b\binom{m-1}{k}$ topics that can be used as a site baseline, and to $b\binom{m-1}{k-1}$ topics that can be used for testing reusability by comparing results on those topics to results on the site baseline topics. In Table 1, for instance, the within-site reusability set for site S_6 includes the first five topics in each subset, e.g. topics numbered $n+1$ through $n+5$ in subset T_1. The within-site baseline includes the first n all-site baseline topics along with the last 10 in each subset, e.g. those numbered $n+6$ through $n+15$ in subset T_1.
2. "between-site": Within each subset T_i, each *pair* of sites contributes to the same $\binom{m-2}{k}$ topics and is held out of the same $\binom{m-2}{k-2}$ topics. The $n + b\binom{m-2}{k}$ total

topics those two sites contribute to form a baseline for comparisons between those sites. The $b\binom{m-2}{k-2}$ topics they were both held out from can be used to determine the between-site reusability. In Table 1, the first topic in each subset can be used for testing reusability between sites S_5 and S_6 against the last six that both contributed to, along with the first n in the baseline.

3. "participant comparison": Within each subset T_i, there are $\binom{m-2}{k-1}$ topics that one site contributes to and another site does not. These topics can be used to evaluate comparing the non-contributing site to the contributing site. In Table 1, if S_5 is the "participant baseline" and S_6 is the "new system", topics numbered $n+2$ through $n+5$ are part of the set used to test reusability.

The values b, n, k are parameters that need to be set by the researchers. Suppose we have an idea of how many total topics (N) will be judged and how many total judgments there will be. This may be based on budget constraints, power analysis, previous work, or most likely a combination of all three. We can express the total number of queries as:

$$N = b\binom{m}{k} + n.$$

Let us further suppose that we want to guarantee that at least n_0 topics are part of the baseline set that all systems contribute to. Then:

$$N \geq b\binom{m}{k} + n_0 \Rightarrow b \leq \frac{N - n_0}{\binom{m}{k}}$$

For a given m and k, we can set

$$b = \lfloor \frac{N - n_0}{\binom{m}{k}} \rfloor \qquad \text{and} \qquad n = N - b\binom{m}{k}$$

Determining k is then a matter of creating a table of values and determining which produces the best distribution of topics among the three types of reusability for answering the questions important to the researchers. Note that larger k provides more topics for between-site experiments, but requires more total topics; smaller k provides more topics for within-site experiments. All design parameters and their relationships to each other are summarized in Table 2.

3.2 Statistical Methods for Analysis

We need to be able to determine whether the evaluation results over "new systems" (restricted to the held-out topics) match the evaluation results over the same systems when they contribute to the judgments. If we were using this design for the TREC Robust track in 2004, for example, we might like to know whether the runs submitted by Johns Hopkins' Applied Physics Lab (APL) are ranked the same when evaluated over 210 topics they contributed judgments to as when evaluated over 39 topics they did not contribute to. This is a statistical question: even if the collection is perfectly reusable, we are evaluating systems over two different sets of topics with two different sample sizes, and we therefore must expect that some evaluation results will change due to chance alone. This must therefore have a statistical answer, i.e. a p-value that will allow us to reject reusability if the evidence is against it.

More specifically, there are three questions of interest:

number of sites	m	fixed by researchers
total number of topics	N	fixed by budget
min. size of baseline set	n_0	fixed by researchers
number of held-out sites	k	variable
number of topic subsets	b	$b = \lfloor (N - n_0)/\binom{m}{k} \rfloor$
size of all-site baseline set	n	$n = N - b\binom{m}{k}$
size of within-site baseline		$n + b\binom{m-1}{k}$
size of between-site baseline		$n + b\binom{m-2}{k}$
size of within-site reuse set		$b\binom{m-1}{k-1}$
size of between-site reuse set		$b\binom{m-2}{k-2}$
size of participant-comparison set		$b\binom{m-2}{k-1}$

Table 2: A summary of parameters of the experimental design and how they relate to each other. Some parameters can be treated as fixed values. At least one is a variable that must be chosen in consideration of certain tradeoffs. The rest are functions of those.

1. Are systems that are significantly different over topics they contributed to also significantly different over topics they did not contribute to? (Likewise with non-significant differences.)
2. Is the relative ordering of systems over topics they contributed to the same as the relative ordering over topics they did not contribute to?
3. Do the system scores averaged over the topics it contributed to match the scores averaged over the topics it did not contribute to?

The first—agreement in statistical significance—is the most important but also the most difficult to discern, so we focus on that. If the first fails, the second—relative orderings being the same—still provides some reusability. The third is a sufficient but not necessary condition for the second; we care about the measures being the same to the extent that they have some extrinsic meaning that we want to keep.

3.2.1 Agreement in statistical significance

Testing for agreement in statistical significance is somewhat complicated. The first step is simple: use some significance test to determine whether pairs of systems are significantly different. We recommend a t-test, possibly adjusting the p-values to account for the family-wise error rate growing with the number of pairwise comparisons (the so-called "multiple comparisons problem" [12]). After performing two sets of pairwise tests (one set for all pairs of systems over the baseline topics, one for the same pairs over the reusability topics), we can form a contingency table showing the agreement in significance between the two sets of tests. The five runs submitted by APL to the TREC 2004 Robust track provide an example:

	baseline tests	
reuse tests	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	6	0
$p' \geq 0.05$	3	1

Among the 10 pairwise comparisons, six resulted in a significant difference being found over both the baseline and reusability topics. Three had a significant difference over the baseline topics but not over the reusability topics. One had no significant difference in either set.

So we can see that there are three errors of omission and none of commission. The question is whether these errors are outside the realm of what is expected. Note that we must expect some errors just because of the difference in topic set sizes between the two experiments. Thus the next step is to construct a contingency table of *expected* agreement between the two sets of tests, then compare our observed values to the expected in a statistically sound way.

We will use *power analysis* to construct the expected contingency table. Power analysis is a very deep topic, and we unfortunately do not have space to go into details. For more information we suggest Cohen's book [10] or two recent papers in the IR literature [9, 16]. The high-level view is that the power of a test is equivalent to the probability that the p-value would be deemed significant for *any* sample of the same size. Power is a function of the *effect size*, the sample size, and the significance level. Effect size is a measure of the degree of difference between two systems over the hypothetical population of topics; for the t-test effect size is estimated as the mean difference in average precisions divided by the standard deviation of the differences. Power monotonically increases with both effect size and sample size.

We can estimate the power of a given pairwise test by estimating the effect size and plugging that along with sample size and significance level (usually 0.05) into a power function (available for most widely-used statistical software packages). This power estimate can then be treated as the expectation that the test would be found significant at the 0.05 level. The power of the comparison over the reusability topics uses the same process, only with the smaller sample size instead of the baseline topic sample size.

For example, the mean difference in average precision between APL runs *rsTs* and *rsDw* is 0.046 over the 210 baseline topics, and the standard deviation is 0.176. The effect size is $0.046/0.176 = 0.260$, which would be considered a moderate effect. The power of a test comparing those two runs over 210 topics is 0.964, i.e. there is a 96% chance that a significant difference between them would be found for any set of 210 topics. If the sample size is reduced to 39 (the size of the reusability set), the power drops to 0.354.

Now the expectation that *both* tests come out significant is simply the product of their estimated powers. For the two runs above, that is $0.964 \cdot 0.354 = 0.341$; we add 0.341 to the number of expected positive agreements. The probability that the first comes out significant but the second does not is $.964 \cdot (1 - 0.354) = 0.623$, so we add that to the number of expected errors of omission. We add $(1 - 0.964) \cdot 0.354 = 0.013$ to the number of expected errors of commission, and $(1 - 0.964) \cdot (1 - 0.354) = 0.023$ to the number of expected negative agreements. Continuing in the same way for all 10 pairs of APL's runs produces the table of expected values:

	baseline expectation	
reuse expect.	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	7.098	0.073
$p' \geq 0.05$	2.043	0.786

By inspection this table is not very different from the observed values. The final step is to verify that statistically. We do that using a χ^2 goodness-of-fit test for whether the observed values match the expected[1]. In this case they do: the p-value of the χ^2 test is 0.88, meaning we cannot conclude that the tables are different, and therefore cannot conclude that the collection is not reusable for this site—we tentatively would say that other sites that are creating runs "like" APL's can trust in the reusability of this collection.

To test between-site reusability, we use the same process, but only test significance between pairs of runs from different sites. For example, if the two sites are APL and IBM, we would only look at significant differences between each APL run and each IBM run (over the intersection of topics they contributed to or were held out from), but not between two APL runs or two IBM runs. Apart from that consideration, the analysis proceeds in exactly the same way. Likewise, participant-comparison uses the same process but uses the topics that one site contributed to and the other did not.

3.2.2 Relative ordering of systems

There are many well-known rank correlation statistics that can be used to determine whether the systems are ordered the same between the two sets of topics. Kendall's τ is the most frequently used; it is calculated by subtracting the number of pairs of systems that have been swapped between two rankings from the number in the same order. Like our significance test procedure above, it calculates counts over pairs of systems; to adapt it to between-site and participant-comparison reusability, we can count pairs that are different between sites while ignoring those from the same site. τ does not have a notion of the expected number of errors that is meaningful for reusability.

Carterette introduced an alternative measure of rank similarity called d_{rank} that takes into account similarity of systems amongst themselves [5]. If the systems are more similar, some reordering is expected, and the measure is smaller. d_{rank} can provide a p-value for the reusability ranking being similar to the baseline ranking.

3.2.3 Agreement in system scores

To determine whether the system scores agree, we can calculate a point estimator such as root mean square error: $RMSE = \sqrt{1/n \sum_{i=1}^{n} (MAP_i - MAP_i')^2}$, where MAP_i is the baseline MAP and MAP_i' is the reusability MAP. The larger this is, the more error is present between the two sets of scores. However, RMSE does not have a known distribution that can be used to determine a p-value, so its interpretation is somewhat subjective.

4. VALIDATION

We present three validation experiments. The first simply demonstrates that it is indeed possible to use our analysis in Section 3.2.1 to disprove reusability. The next two show conversely that if a collection is reusable the p-value is not likely to be low.

4.1 Disproving reusability

A very simple way to validate that reusability will be rejected when it is not true is to simulate evaluation over a non-reusable collection. For example, we can use random number generation to simulate evaluation measures for m systems and show that the χ^2 p-value will be low when the simulation is explicitly set up so that the evaluation measures differ between the baseline and reusability sets. We drew measures from beta distributions (ensuring they would be between 0 and 1) such that the measures drawn for reusability topics for one run would be lower than those

[1]In this case, because the number of observations is small, we actually use a randomized "exact" version of the χ^2 test.

	baseline	
reuse	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	$O = 196$ $E = 189.5$	$O = 2$ $E = 4.3$
$p' \geq 0.05$	$O = 57$ $E = 62.1$	$O = 45$ $E = 44.1$

Table 3: Observed versus expected agreement in significance results for within-site reusability aggregated over all Million Query 2008 sites. The χ^2 p-value is 0.58, indicating no evidence to reject reusability.

drawn for the baseline topics for the same run and as a result the significance tests involving those runs would not agree. The result is that as the number of reusability topics increases, the p-values decrease, with 50 reusability topics in this scenario producing a p-value less than 0.01.

4.2 Robustness to differences in topic samples

By robustness to differences in topic samples, we mean that conclusions about reusability are not expected to be confounded by the fact that the tests are based on different size samples of different topics (as in Section 3.2.1 above). To show this, we set up an idealized scenario in which the test collection *must* be reusable and show that our analysis will not reject reusability.

Our data is the 2008 TREC Million Query (MQ) track data consisting of 564 topics with 15,000 total judgments collected from 25 systems submitted by 9 different sites [1]. Every run contributed judgments to every topic; none were held out. We chose $n_0 = 200$ topics to be the baseline that all systems "contribute" to. For each of the remaining 364 topics, we "held out" $k = 2$ sites. Plugging into the formula above results in $b = 10$ topic sets.

The AP value for each system/topic is simply that calculated for the track. This makes a 100% reusable collection: the AP estimates on the "held-out" topics are exactly the same as they were when the systems actually contributed judgments. This is (intentionally) highly artificial, but note that it is not meant to be a simulation of judgment collection or of evaluation. It is a boundary case to demonstrate that our conclusions will not be confounded by variance in the topic samples *when reusability is true*. There are other sources of bias and variance that this test does not address.

Rather than apply the procedure described in Section 3.2.1 to each site individually (running the risk of multiple comparisons problem), we aggregated the contingency tables and expected contingency tables across sites to obtain two tables representing all within-site comparisons. They are shown together in Table 3. The χ^2 p-value is 0.58, indicating no evidence to suggest the difference in topic samples is causing a problem. We did the same for between-site reusability and participant-comparison reusability; the χ^2 p-values are 0.54 and 0.36, respectively.

4.3 Robustness to held-out systems

Another possibility is that holding certain systems out will inject bias into topic evaluations. For example, if a very good system that retrieves many relevant documents is held out, evaluation results for the other systems may not be as accurate, even when reusability holds in other cases. To test this we use simulation in the Robust 2004 data described

	baseline	
reuse	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	$O = 130$ $E = 135.4$	$O = 17$ $E = 13.9$
$p' \geq 0.05$	$O = 127$ $E = 121.6$	$O = 160$ $E = 163.1$

Table 4: Observed versus expected agreement in significance results for within-site reusability aggregated over all Robust 2004 sites. The χ^2 p-value is 0.74, indicating no evidence to reject reusability.

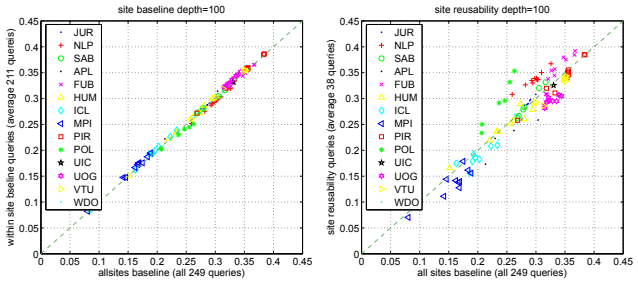

Figure 1: Robust 2004 simulation with 2 sites held out per topic. Judgments were based on a pool of depth 100. The left plot compares MAP over the 210 site baseline topics to true MAP calculated with all judgments. The right compares MAP over the 39 site reusability topics to true MAP.

above. Note that Robust 2004, despite being large by TREC standards, is fairly small for this design; because there are only 249 topics, we have no all-site baseline set, and we cannot hold more than $k = 2$ sites out. The number of topics we have for each of the three types of tests is limited (39 for within-site, only 3 for between-site). The advantage is that Robust 2004 has very many relevance judgments, so we can simulate pools of any depth.

Once again, this is validation that the design works when reusability is true. To ensure reusability to the greatest degree possible, we simulated a depth-100 pool. That is, for each topic, runs submitted by two sites were held out of simulated judging; the other 12 sites had their top 100 ranked documents judged according to the existing judgments in the TREC qrels file. We then evaluated all runs using that pool and separated them into systems that contributed and systems that were held out.

We only have enough topics for within-site analysis. The observed and expected significance results are shown in Table 4; the p-value is 0.74, indicating no evidence to reject reusability. We performed the same test on shallower pools; for pools of depth 10, 20, and 50, the p-values are 0.63,0.58, and 0.60, respectively. Figure 1 shows the comparison of evaluation results on different topic sets in the depth-100 pool. Note that the fact that we have only 39 topics for reusability testing is somewhat limiting, however.

5. IN SITU REUSABILITY EXPERIMENT

The analysis above provides evidence that our design is correct. We next observe it in a real experimental setting: judgment collection for the 2009 TREC Million Query (MQ) track [8]. Eight participating sites submitted a total of 35

runs over 1,000 queries. The corpus was the Category B subset of the new ClueWeb09 web collection. 638 of the 1,000 queries were converted to full topics and judged; of those, 146 formed the all-site baseline to which every run contributed judgments. The remaining 492 topics had two sites held out during judging. Held-out sites were selected by round-robin scheduling. Assessors did not know whether they were judging a reusability topic or not, and topic order was randomized, so there is no reason to suppose that the reusability topic sample is biased compared to the baseline sample. Assessors made a total of 34,534 judgments (54 per topic on average), of which 26% were either relevant or highly relevant. There were 95 topics for which no relevant documents were found.

The Million Query track uses two official evaluation measures, statMAP and MTC's "expected" MAP. Both are estimates of average precision, but they are designed for different purposes. statMAP is an unbiased estimator of average precision. MTC EMAP is a biased estimator meant to provide good comparative evaluation.

Our goal in this section is to determine the extent to which this test collection is reusable.

5.1 Results

Reusability results for MQ are illustrated in Figure 2, which shows statMAP (top) and MTC EMAP (bottom) scores of runs over (a) 145 baseline against 170 site baseline topics (left), (b) 170 site baseline against 160 site reuse topics (center), and (c) 145 baseline against 160 site reuse topics (right). Each run was evaluated over all the topics it contributed to and all the topics it was held out from, but since different sites contributed to different topics, no two sites were evaluated over exactly the same set of topics.

As mentioned in Section 4, differences in mean scores over baseline topics and mean scores over reusability topics for a given site may be due to a number of different effects: (1) the baseline and reuse topics are two different topic sets of different size; (2) apart from the site under study there are two other sites that did not contribute documents to each reusability topic; (3) the site under study itself did not contribute documents to the reuse topics (this is the actual effect we would like to quantify); and finally, (4) for this particular study the fact that both methods evaluate runs with a very small number of documents introduces some variability even in the baseline topics.

The plots in Figure 2 attempt to separate the second and third effects. Essentially, the comparison of the mean scores between the 145 baseline topics and the 160 site reuse topics (right) summarizes the results of the reusability experiment, and it is what an actual new site would observe by using the MQ 2009 collection. StatMAP scores over the reuse topics are positively correlated with the statMAP scores over the baseline topics, though the correlation is rather weak. MTC EMAP scores over these two sets of topics are well correlated. One can consider the other two plots as the decomposition of the effects seen in the right plot. The left plot illustrates the effect of holding out sites other than the site under study. For the statMAP case this has a rather strong effect on the scores computed, though it is minimal for the MTC scores. The middle plots try to isolate the effect of holding out the site under study. As can be seen, this also has a strong effect on the statMAP scores, while the effect is mild in the case of the MTC scores.

	baseline	
reuse	$p < 0.05$	$p \geq 0.05$
$p' < 0.05$	$O = 257$	$O = 41$
	$E = 302.5$	$E = 26.2$
$p' \geq 0.05$	$O = 133$	$O = 100$
	$E = 85.1$	$E = 117.2$

Table 5: Observed versus expected agreement in significance results for between-site reusability aggregated over all Million Query 2009 sites. The χ^2 p-value is 0, indicating sufficient evidence to reject reusability.

The plots give a visual sense of reusability, suggesting within-site may be acceptable at the level of rank agreement if not score agreement, but between-site is likely not acceptable. To quantify this, we computed three correlation statistics as described in Section 3.2.2. First we computed the overall Kendall's *tau* between the ranking induced by the scores in the two topic sets. This is a rough estimate of the between-site reusability. For statMAP scores this is 0.7643, while for MTC EMAP scores this is 0.8350, both of which are rather low. Next we computed the Kendall's τ among the runs of each individual site to estimate within-site reusability; Table 6 shows these. Note that the values are not comparable across sites since the number of runs compared affects the Kendall's τ values. Finally, we computed a τ-like correlation to quantify the ability to compare "new" runs to contributing participants. For each site, we count the number of its reusability runs that are correctly ordered against the baseline runs and the number that have been swapped with a baseline run. Every comparison involves exactly one run for that site; for this measure we do not compare two runs from the same site or two runs from a different site. The final value is determined identically to Kendall's τ; the set of values can be seen in Table 6.

The significance test agreement procedure, when applied to this data, suggests that there is not enough evidence to reject within-site reusability ($p > 0.5$), but there is more than enough to reject between-site reusability ($p < 0.01$). To explain how within-site reusability holds despite some of the low τ correlations in Table 6, we note that τ is not able to capture anything about whether swaps are "reasonable". The lowest τ is -0.6 for UIUC, but by inspection (Fig. 2) UIUC's systems are all very close to each other. It is perfectly reasonable that they would be ordered differently over another set of topics, and thus the low τ is not a concern. For between-site reusability, however, we have seen that it is unlikely; that the χ^2 test confirms this is a point in its favor. The full contingency table for between-site reusability is shown in Table 5.

6. CONCLUSIONS

We have proposed an experimental design that can be used during construction of large test collections to collect evidence for or against the future reusability of the collection. It is appropriate for when the set of judgments is too small to be able to evaluate reusability through simulation; since test collections are moving in this direction, some framework will be necessary for determining whether these collections can be reused. We presented tools for statistical analysis and demonstrated their use in artificial data, a

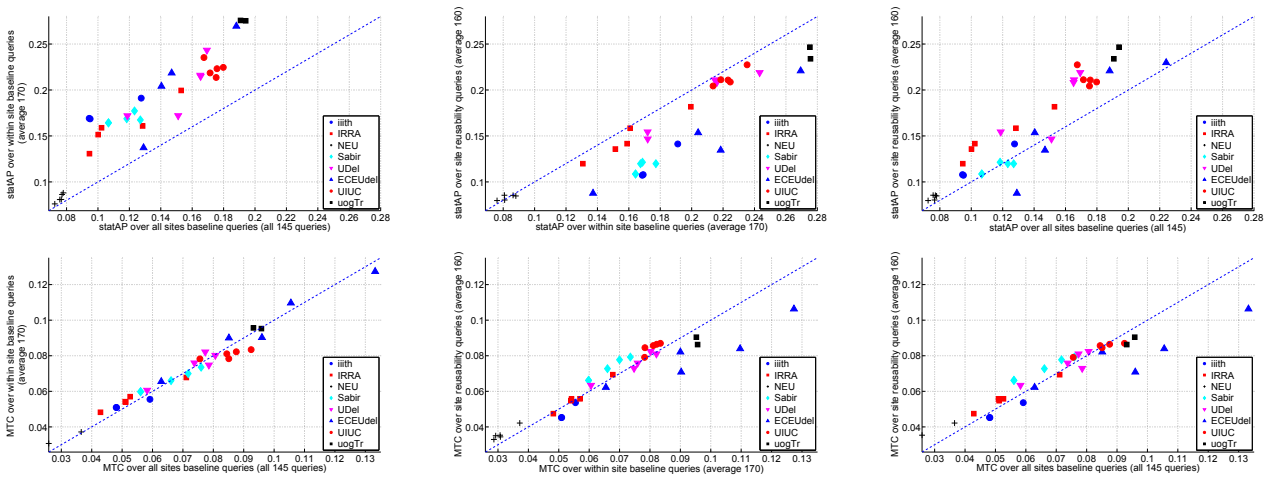

Figure 2: StatMAP and MTC EMAP scores of systems over (a) 145 baseline against 170 site baseline topics, (b) 170 site baseline against 160 site reuse topics, and (c) 145 baseline against 160 site reuse topics.

		iiith	IRRA	NEU	Sabir	UDel	ECEUdel	UIUC	uogTr
within-site τ	statAP	0.333	1.000	0.200	0.333	0.800	0.800	-0.600	1.000
	MTC	0.333	0.800	1.000	1.000	0.600	0.800	0.800	1.000
participant comparison	statAP	0.750	0.547	1.000	0.987	0.573	0.773	0.773	0.939
	MTC	0.938	1.000	1.000	0.840	0.933	0.707	0.947	0.909

Table 6: Rank correlations based on Kendall's τ for site baseline to site reusability (top) and for comparison of site reusability to the "original" TREC runs excluding those treated as new (bottom).

simulation experiment, and a real-life implementation of the design; in general their results confirm our intuitions about the evaluation.

Clearly there is much more and much deeper analysis we could do. For this work we chose to present some of the topics we felt were most important in presenting this methodology, but we certainly intend to continue investigating other tools for analysis, more sophisticated statistical methods, and of course IR-centric implications for the failure (or lack of failure) of reusability when it happens.

Acknowledgements

The authors gratefully acknowledge support by the European Commission who funded parts of this research within the Accurat project (FP7-ICT-248347) and by the Marie Curie IIF (FP7-PEOPLE-2009-IIF-254562).

7. REFERENCES

[1] J. Allan, J. A. Aslam, B. Carterette, V. Pavlu, and E. Kanoulas. Overview of the TREC 2008 million query track. In *Proceedings of TREC*, 2008.

[2] J. A. Aslam and V. Pavlu. A practical sampling strategy for efficient retrieval evaluat ion, technical report.

[3] S. Büttcher, C. Clarke, P. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proceedings of SIGIR*, pages 63–70, 2007.

[4] B. Carterette. Robust test collections for retrieval evaluation. In *Proceedings of SIGIR*, pages 55–62, 2007.

[5] B. Carterette. On rank correlation and the distance between rankings. In *Proceedings of SIGIR*, 2009.

[6] B. Carterette, J. Allan, and R. K. Sitaraman. Minimal test collections for retrieval evaluation. In *Proceedings of SIGIR*, pages 268–275, 2006.

[7] B. Carterette, E. Gabrilovitch, V. Josifovsky, and D. Metzler. Measuring the reusability of test collections. In *Proceedings of WSDM*, 2009.

[8] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. Overview of the TREC 2009 million query track. In *Notebook Proceedings of TREC*, 2009.

[9] B. Carterette and M. D. Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of CIKM*, pages 643–652, 2007.

[10] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum, 2nd edition, 1998.

[11] D. Harman. Overview of the second text retrieval conference (trec-2). *Inf. Process. Manage.*, 31(3):271–289, 1995.

[12] J. P. A. Ionnidis. Why most published research findings are false. *PLoS Med.*, 2(8), 2005.

[13] T. Sakai. Comparing metrics across trec and ntcir: the robustness to system bias. In *Proceedings of CIKM*, pages 581–590, 2008.

[14] E. M. Voorhees. The philosophy of information retrieval evaluation. In *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, pages 355–370, London, UK, 2002. Springer-Verlag.

[15] E. M. Voorhees. Overview of trec 2009. In *Proceedings of TREC*, 2009. Notebook draft.

[16] W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *Proceedings of CIKM*, pages 571–580, 2008.

[17] J. Zobel. How Reliable are the Results of Large-Scale Information Retrieval Experiments? In *Proceedings of SIGIR*, pages 307–314, 1998.