Statistical comparisons of non-deterministic IR systems using two dimensional variance

Gaya K. Jayasinghe

RMIT University, Australia

William Webber

William Webber Consulting, Australia

Mark Sanderson

RMIT University, Australia

Lasitha S. Dharmasena

Deakin University, Australia

J. Shane Culpepper

RMIT University, Australia

Abstract

Retrieval systems with non-deterministic output are widely used in information retrieval. Common examples include sampling, approximation algorithms, or interactive user input. The effectiveness of such systems differs not just for different topics, but also for different instances of the system. The inherent variance presents a dilemma – What is the best way to measure the effectiveness of a non-deterministic IR system? Existing approaches to IR evaluation do not consider this problem, or the potential impact on statistical significance. In this paper, we explore how such variance can affect system comparisons, and propose an evaluation framework and methodologies capable of doing this comparison.

Using the context of distributed information retrieval as a case study for our investigation, we show that the approaches provide a consistent and reliable methodology to compare the effectiveness of a non-deterministic system with a deterministic or another non-deterministic system. In addition, we present a statistical best-practice that can be used to safely show how a non-deterministic IR system has equivalent effectiveness to another IR system, and how to avoid the common pitfall of misusing a lack of significance as a proof that two systems have equivalent effectiveness.

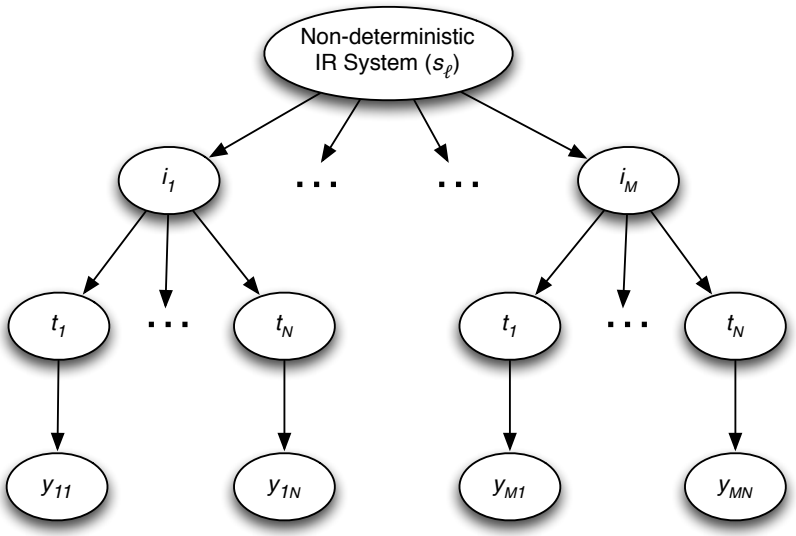
Keywords: Information Retrieval Evaluation, Non-determinism, Randomisation, Statistical Analysis, Distributed IR, Personalised Web Search

1. Introduction

The output of IR systems can be non-deterministic. This can be due to the use of sampling in randomized algorithms [2, 7, 6, 20, 31, 32, 33, 30, 34], or training based on user input [17, 22, 23]. Evaluating the effectiveness of

Email addresses: gaya.jayasinghe@rmit.edu.au (Gaya K. Jayasinghe), william@williamwebber.com (William Webber), mark.sanderson@rmit.edu.au (Mark Sanderson), lasitha.dharmasena@deakin.edu.au (Lasitha S. Dharmasena), shane.culpepper@rmit.edu.au (J. Shane Culpepper)

these systems can be difficult because each run can produce a different IR system instance, hence different results and effectiveness scores. Figure 1 illustrates this situation. How can the effectiveness of an IR system be measured if each run can produce a different output?



s_ℓ Non-deterministic IR system ℓ,
i_m IR system instance m produced using non-deterministic IR system ℓ, where $m = 1, ..., M$
t_n $n{\cdot}th$ topic for evaluation, where $n = 1, ..., N$.
y_{mn} Effectiveness for topic n on IR system instance m.

Figure 1: Variation in effectiveness for a non-deterministic IR system along two dimensions (IR system instances and topics), with effectiveness for topics grouped within IR system instances.

An evaluation based on a single instance of an IR system may produce results which can change an experimental conclusion of whether one IR system has better effectiveness than another. The obvious solution is to generate many system instances and make statistically grounded inferences about the overall average effectiveness.

Experiments in IR add another layer of complexity to the problem as retrieval effectiveness varies by topic [35]. Recall that the effectiveness of an IR system instance is characterized and compared using average effectiveness under whatever evaluation metric is employed across a set of topics. Differences are tested for statistical significance across a hypothetical population of topics using a significance test. However, standard significance tests only support one source of variability, in this instance the choice of topics. The use of non-deterministic IR systems introduces an additional dimension of variability. So, how can we determine that one IR system is significantly better than another when averaging across all topics and all possible IR system instantiations?

In order to understand the uncertainty due to variability in more than one dimension, consider the following simulation. We consider the effectiveness of each topic on each non-deterministic IR system instance as composed of two effects. First, we randomly sample the "topical effect" between 0 and 1 from a uniform distribution for each topic. Next, we realize the effect due to non-determinism in the system capped between 0 and 1 from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. Now assume that the simulated effectiveness for each topic on each system instance is the *Euclidean distance (L_2 norm)*[1] between the two effects, normalized to a value between 0 and 1. Consider another synthetic system with a deterministic output. We randomly sample the effectiveness between 0 and 1 for each topic from a uniform distribution for the deterministic system. For the simulation, we sample two sets of 50 topical effects, one for each system being compared and 100 system instances for a given μ and σ from the respective distributions. We then compare the effectiveness for each system instance with the deterministic system using a standard t-test, and

[1]The Euclidean distance (L_2 norm) is the square root of the sum of the squares of each effect.

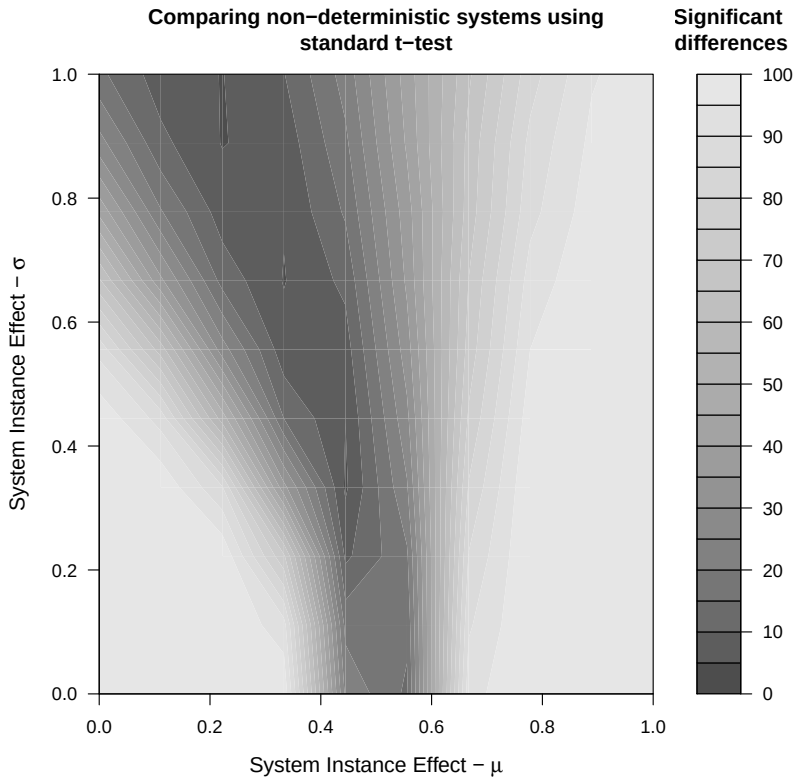


Figure 2: Results for comparing simulated non-deterministic IR system instances with a deterministic IR system. The system instance effect is randomly sampled from a normal distribution for each IR system instance. The topic effect is randomly sampled from a uniform distribution between 0 and 1. The effectiveness for a topic-system instance pair is the Euclidean distance of respective component effects normalized to a value between 0 and 1. Each system instance is compared with the deterministic system having only a topic effect using a standard paired t-test, and the significant differences are counted at a $p < 0.05$.

note the mean of the number of significant differences over 100 repetitions of the above process as μ and σ are varied. This is equivalent to repeatedly comparing an instance of a non-deterministic IR system having two dimensional variance with a deterministic IR system. The results are illustrated in Figure 2. The expected effectiveness for the deterministic system is 0.5. Similarly, the expected effectiveness for the non-deterministic IR system is 0.5 when μ is 0.5. Therefore, the majority of comparisons show no significant difference when μ is 0.5 with a small σ. When the standard deviation is increased, the majority of the comparisons with a μ less than 0.5 are not significantly different. However, a comprehensive analysis of the impact on a standard significance test due to variation in system instance effect for a non-deterministic IR system is not possible with the above simple simulation. The take home message from the simulation is that all comparisons derive the same conclusion when number of significant differences are 0 (darkest gray) or 100 (lightest gray). An intermediate shade of gray implies that conclusions derived using some instances contradict the conclusions from the rest of the instances when using the same non-deterministic IR system. For example, 50% of the comparisons show no significant difference at the center of the gray scale, while the rest imply the opposite. So the gradual hill climbing pattern (from lightest gray to darkest gray) of the graph implies that experimental conclusions derived with certain system instances can be contradicted by the others and exemplifies the pitfalls in using standard significance tests to compare non-deterministic IR systems.

In this paper, we extend our prior work on non-deterministic system evaluation [19]. In our previous work, we used a linear model to evaluate non-deterministic IR systems, and a limited case study using a single evaluation metric (NDCG@10) on the TREC GOV2 dataset was presented. In this work, we include an alternative evaluation methodology based on the bootstrap, and show that the two evaluation methodologies produce similar conclusions. We

also demonstrate the applicability of proposed solutions using a simulation in addition to the original case study. We also extend coverage to include more evaluation metrics (NDCG@10, P@10, MAP) and two different test collections (TREC GOV2, and ClueWeb '09B).

Our Contribution:

1. We propose a pair of methodologies, bootstrapping (Section 3.1.1) and multivariate linear modelling (Section 3.1.2) to solve the two-dimensional significance testing problem of comparing a non-deterministic IR system with a deterministic IR system. Both methods result in equivalent inferences (Section 4 and Section 5.2).
2. We extend the multivariate linear modelling approach to support the comparison of two non-deterministic IR systems (Section 3.2).
3. We verify the applicability of the solutions in real world scenarios using a case study of common sampling algorithms – shard construction and centralized resource allocation in distributed IR [31, 20] (Section 5).
4. We present a methodology for comparing a non-deterministic IR system for equivalent or greater effectiveness with either a deterministic or another non-deterministic IR system using a two dimensional evaluation (Section 5.4).

2. Background

Since the early 1960s, researchers in the IR community have benefited greatly from sharing test collections. A collection typically is composed of documents, test topics, and relevance judgments. Using these collections, IR systems can be compared by calculating the effectiveness of each system using a common metric such as MAP or NDCG. However, proving that one system is "better" than any other simply on the basis of achieving a higher effectiveness score on a collection is not as straightforward as it might initially seem. The subtle differences in average score might just be coincidental to the particular set of topics selected, and may not hold across the full population of possible queries and topics. In response, IR research has adopted the practice of hypothesis testing.

2.1. Hypothesis testing

Let a_n and b_n be the effectiveness measured using an evaluation metric for topic t_n on System $\mathcal{A}$ and System $\mathcal{B}$ for a random sample of N topics. Given that System $\mathcal{A}$ has achieved a higher average effectiveness score than System $\mathcal{B}$ on the sample of topics, the hypothesis test determines the confidence that a similar behaviour is observed across the population of all topics. This question is answered by setting two hypotheses: H_0, called the null hypothesis, assumes the two systems have equal effectiveness across the population of topics; H_1, called the alternative hypothesis, assumes they are not equal.

$$H_0 : \mathcal{M}_{\mathcal{A}} = \mathcal{M}_{\mathcal{B}} \tag{1}$$

$$H_1 : \mathcal{M}_{\mathcal{A}} \neq \mathcal{M}_{\mathcal{B}}. \tag{2}$$

Here, $\mathcal{M}$ is the average effectiveness for the population of topics. The question then is: If the null hypothesis H_0 were true (that is, the two systems have hypothetically equal mean effectiveness), how likely would the difference in effectiveness observed in the sample be due to the chance of random sampling? The likelihood of two systems being hypothetically equal is known as the p-value of the test. Precisely how the p-value is calculated differs from test to test, but the basic idea is to compare the observed average difference (Δ) against the observed variability in per-topic score differences. A p-value below a commonly accepted level of significance (α) implies a systematic difference that cannot be attributed purely to the chance in selection of topics. So, the null hypothesis is rejected and the alternative hypothesis is accepted. Conversely, a larger p-value fails to reject the null hypothesis and therefore the difference is not statistically significant. However, failing to reject the null hypothesis is not the same as accepting it. Common levels of significance are 0.1, 0.05, and 0.01.

Commonly used approaches for hypothesis testing of IR systems include *Student's paired t-test*, the *bootstrap test* [14, 15, 16, 29, 26], the multivariate linear model test [9], the *sign test*, the *Wilcoxon signed rank test* [37], and *Fisher's randomization test* [18]. Next, we summarize the use of these approaches.

ALGORITHM 1: Bootstrap algorithm for comparing two deterministic IR systems.

$BS \leftarrow$ Number of bootstrap samples;
$N \leftarrow$ Number of topics;
$a_n \leftarrow$ Vector of scores for IR System $\mathcal{A}$ on the $n{\cdot}th$ topic, where $n = 1, ..., N$;
$b_n \leftarrow$ Vector of scores for IR System $\mathcal{B}$ on the $n{\cdot}th$ topic, where $n = 1, ..., N$;

// Compute $t(z)$.
$z \leftarrow [(a_1 - b_1), ..., (a_N - b_N)]$;
$t(z) \leftarrow \overline{z} / \sqrt{\hat{\sigma}_z^2 / N}$;

// Compute p-value.
$count \leftarrow 0$;
$bootstrap\_samples \leftarrow []$;
$total\_shift \leftarrow 0$;
for $i = 1$ *to* BS **do**
  $z_i^* \leftarrow$ Resample N items with replacement from z;
  $bootstrap\_samples.append(z_i^*)$;
  $total\_shift \leftarrow total\_shift + mean(z_i^*)$;
end
$shift \leftarrow total\_shift / BS$;
for $i = 1$ *to* BS **do**
  $z_i^* \leftarrow bootstrap\_samples[i]$;
  for $n = 1$ *to* N **do**
    $z_i^*[n] \leftarrow z_i^*[n] - shift$;
  end
  $t(z_i^*) \leftarrow \overline{z}_i^* / \sqrt{\hat{\sigma}_{z_i^*}^2 / N}$;
  if $|t(z_i^*)| \geq |t(z)|$ **then**
    $count \leftarrow count + 1$;
  end
end
$p\_value \leftarrow count / BS$;

2.1.1. Student's paired t-test

Recall that a significance test compares the difference in effectiveness between two IR systems against the variability in per-topic score differences. This is achieved with a Student's paired t-test as follows. Let $z_n = (a_n - b_n)$ be the difference in effectiveness for the $n{\cdot}th$ topic between System $\mathcal{A}$ and System $\mathcal{B}$, and $z = [z_n]$. A student's paired t-test computes a *test statistic* as follows:

$$t(z) = \frac{\overline{z}}{\sqrt{\hat{\sigma}_z^2 / N}}, \tag{3}$$

where $\overline{z}$ is the mean of z, and $\hat{\sigma}_z^2$ is the variance estimated as follows:

$$\hat{\sigma}_z^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (z_i - \overline{z})^2. \tag{4}$$

The test statistic can be used to derive the p-value from the t distribution with $N - 1$ degrees of freedom.

2.1.2. Bootstrap test

An alternative is to empirically estimate the p-value using evidence gathered from a sample. A bootstrap approach for comparing two IR systems ($\mathcal{A}$ and $\mathcal{B}$) for mean effectiveness with pairing by topic is shown in Algorithm 1. The

test creates *BS* bootstrap samples of size similar to the original sample by repeatedly resampling with replacement. The bootstrap samples are then recentered by deducting the mean of bootstrap samples from each value in the bootstrap samples. So, the recentered samples can be assumed to be from two equivalently effective systems. In prior literature bootstrap samples have been recentered by deducting the mean of the original sample from each element in bootstrap samples [26]. However, this approach could produce a bias equal to the difference between the mean of the mean estimates derived from bootstrap samples and the single mean estimate obtained from the original sample. Therefore, the bootstrap samples are recentered in the manner shown in Algorithm 1 to obtain a bias-corrected estimate of the statistic. Now the evidence for the null hypothesis can be observed by comparing individual recentered bootstrap samples with the original sample. Let $t(z)$ and $t(z_i^*)$ be the t-statistic for the original and the $i{\cdot}th$ recentered bootstrap samples. (Use of the t statistic makes this known as a *Studentized bootstrap*.) Given a probability function p, the p-value is computed as:

$$p(|t(z^*)| \geq |t(z)|). \tag{5}$$

2.1.3. Multivariate Linear Model Test

Another approach for testing significance is through linear modelling. When comparing two IR systems, the following linear regression model is applied:

$$y_{\ell n} = \gamma + s_\ell + t_n + \varepsilon_{\ell n}. \tag{6}$$

Here, $y_{\ell n}$ is the effectiveness of system ℓ on topic n and variance is assumed to come solely from the selection of topics. The effectiveness due to the two factors or effects is modelled as: the system effect s_ℓ (the effectiveness of one system relative to the other system) which has two levels for systems $\mathcal{A}$ and $\mathcal{B}$, and the topic effect t_n which has N levels, one for each topic (the difficulty of the topic relative to all other topics). The intercept in the model, γ, is the average effectiveness of all systems across all topics, and $\varepsilon_{\ell n}$ is the residual or error term that accounts for the deviation of the observed effectiveness. The p-value for the system effect can either be computed using the t-statistic derived from $\Delta_{s_\ell} / \sqrt{\sigma^2/N}$ or from ANOVA which follows a Student's t distribution with $N - 1$ degrees of freedom. Here Δ_{s_ℓ} is the difference between the two system effects, σ^2 is the variance of the residuals, and N is the number of topics.

Alternatively, the same can be modeled using a *linear mixed effect (LME)* model, consisting of fixed (non-random) and random effects [10, 25]. For this scenario, sampled topics produce a random effect, and systems produce a fixed effect. As we will see later, LME can be used to build complex models that capture repeated measurements and hierarchical grouping. For large samples, the p-value can be computed from a t-statistic obtained from LME with $N - f$ degrees of freedom, where f is the number of fixed effect parameters ($f = 1$ for the above scenario) [4].

Another way to compute the p-value is to use Bayesian inference and *Markov Chain Monte Carlo (MCMC)* simulations [4]. MCMC sampling can be derived from the posterior probabilities of the parameter subsets of the LME model (σ, parameters defining the variance-covariance for random effects, and fixed and random effects). The posterior probability distributions are produced with repeated and cyclic sampling from each parameter subset conditioned on the other two parameter subsets and on the data, thus making variance of all other parameter subsets reflect the variance for each parameter subset. For this process of sampling a Markov chain is used, so that the sampling spends more effort on regions of most importance [3]. The posterior distribution of the fixed system effect factor is expected to follow at least a near symmetric normal distribution which can be used to compute the p-value. Deriving the posterior distribution also allows us to obtain the highest posterior density (HPD) interval which is analogous to the standard confidence interval. A ϑ % HPD interval represents the shortest interval enclosing $(1 - \vartheta)$ % of the posterior probability mass of the distribution. Therefore, the HPD interval is considered a better representation than a standard error interval [11].

All of the significance testing approaches discussed so far assume an IR system with a deterministic output, where only one observation exists per topic. Carterette et al. [10] and Robertson and Kanoulas [25] observed that the measured effectiveness for the same topic on the same IR system can vary under varying experimental conditions leading to repeated observations along a single dimension.

2.1.4. Repeated Measurements

Topical variance can be thought of as combining two components: first, measurement or model error; and second, the fact that some systems do better on some topics than others (both systems and topics). If we only have one

Practise	Null hypothesis (H_0)	Desired outcome	Inference	Encouraged to
Incorrect	$\mathcal{M}_\mathcal{A} = \mathcal{M}_\mathcal{B}$	Accept H_0	Not "statistically different"	Reduce sample
Correct	$\lvert\mathcal{M}_\mathcal{A} - \mathcal{M}_\mathcal{B}\rvert \geq \delta$	Reject H_0	Difference is significantly less than consequential	Expand sample

Table 1: Incorrect and correct practises of testing for "statistically significant equivalence" of systems $\mathcal{A}$ and $\mathcal{B}$.

observation per topic, we cannot separate these two factors; but if we have repeated observations, the two factors can be separately estimated. Taking system effect as fixed, and topic and system–topic interaction effects as random, the above can be modeled with LME, as follows:

$$y_{\ell ni} = \gamma + s_\ell + t_n + st_{\ell n} + \varepsilon_{\ell ni}. \tag{7}$$

Here $y_{\ell ni}$ is the effectiveness on the $i{\cdot}th$ observation of system ℓ on topic n, $st_{\ell n}$ is the system-topic interaction effect, and $\varepsilon_{\ell ni}$ represents the (random) error of a single observation. However, the model only makes sense if different observations on the same system-topic pair lead to different scores. In Robertson and Kanoulas [25], this variability in system-topic scores is observed over different document sets, whereas in Carterette et al. [10] the variability is in different user types. Our focus in this paper is on the added dimension of variability introduced by using a sampling-based algorithm or user feedback. Hence, variability exists in two dimensions, IR system instances and topics, with topical variation grouped within the IR system instances.

Statistical hypothesis testing only minimises the uncertainty associated with a scientific conclusion, and has been surrounded by controversy for many years [12]. The main criticism is that p-values can be misinterpreted. For example, one may incorrectly consider failing to accept the alternative hypothesis as accepting null hypothesis. This potential problem is discussed in detail in Section 2.3.

An alternative solution relying on p-values is to use confidence intervals. A confidence interval for the differences in effectiveness for a set of topics that does not include zero implies a systematic difference between two systems at a predetermined level of confidence $(1 - \alpha)$. Assessing the probability of two systems as being equal is not appropriate, as the performance of two systems will always have some differences. For example, the effectiveness of two IR systems will always be significantly different when two systems are compared across a huge number of topics. Hence, statistical significance should be reported with *effect size*, which quantifies the magnitude of the difference in average effectiveness.

2.2. Effect Size

The effect size [13] can be computed as:

$$\text{effect size} = \frac{\bar{z}}{\hat{\sigma}_z}. \tag{8}$$

The common rule of thumb when classifying effect size is: 0.8 and above is a large effect, 0.5 to 0.8 is a medium effect, and 0.2 and 0.5 is a small effect. An effect of less than 0.2 is considered negligible.

2.3. Evaluating Systems for Similarity

Systems are often required to be compared for similar effectiveness when sampling or approximation is used to increase efficiency. This is done to ensure that the efficiency gains do not harm the overall effectiveness of the system being evaluated. Hypothesis testing as described above cannot be used for such a comparison. In reality, the null hypothesis is a statistical straw man. Even when two systems are identical, using the failure of a statistical significance test to reject the null hypothesis as a proof for equivalence promotes bad practice. If the main goal is to find no significant difference between two systems, an experimentalist might be tempted to reduce the sample size. This minimizes the likelihood of finding any significant difference between the baseline and the new system and "justifies" the experimental objective. This is clearly the wrong approach.

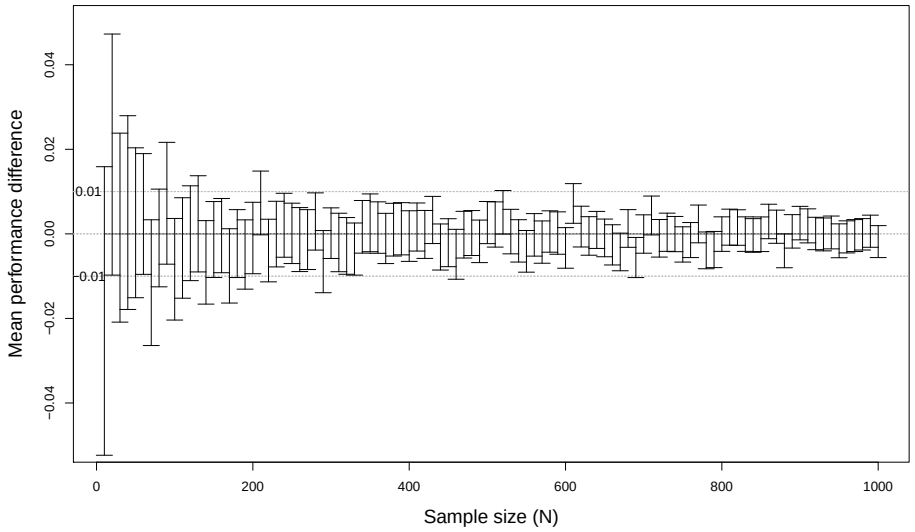


Figure 3: The 95% confidence interval for mean performance difference between two equal systems $\mathcal{A}$ and $\mathcal{B}$ with increasing sample sizes.

The conclusion that two systems are equivalent can only be reached after seeing the population. Therefore, determining the (rough) similarity of the effectiveness of two systems using the performance observed for a sample of experimental units is not as straightforward as it might initially seem.

Inevitably, an average estimated on a sample is subject to variance, and hence uncertainty. Therefore, if hypothesis testing is used to recognize "statistically significant equivalence", a better practice is to consider the least difference in mean performance, δ, that one would consider as being consequential [1, 36]. Here, $|\mathcal{M}_{\mathcal{A}} - \mathcal{M}_{\mathcal{B}}| \geq \delta$, becomes the null hypothesis; and $|\mathcal{M}_{\mathcal{A}} - \mathcal{M}_{\mathcal{B}}| < \delta$ is the alternative hypothesis. The null hypothesis, that is, is that the difference between the two systems (on the full population of topics) is consequential; if we reject this null hypothesis, we instead accept the alternative hypothesis, that any difference between the systems is inconsequential; that is, that they are approximately equal in their effectiveness. This process not only makes statistical sense, it also drives good practice – the experimentalist is encouraged to increase the sample size and thus the accuracy of the average estimates. The pros and cons of these two methods of testing for statistically significant equivalence are summarized in Table 1.

The above principle can also be indirectly tested using confidence intervals. If the confidence interval is above $-\delta$ and below δ, one can conclude that any difference between the two systems are not more than δ at a given level of confidence. Hence, one method is not consequentially different than the other method.

In order to demonstrate the approach, we now present a simple simulation as a concrete example. Consider the case, where the mean effectiveness of $\mathcal{A}$ and $\mathcal{B}$ are equal, such that the expectation of $\mathcal{M}_{\mathcal{A}} - \mathcal{M}_{\mathcal{B}}$ is 0. Assume that the difference in effectiveness between the two systems for a synthetic set of topics is sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$. Here the choice of δ is defined by the user, and is problem dependent. The value should be set to the minimum difference in effectiveness that is considered consequential. Recall that a lower value of α is preferred to show difference in effectiveness between two systems. Similarly a lower value of δ is desired to demonstrate similarity between two systems. Although, potential values for α are commonly accepted, no convention exists for δ. The most appropriate value of δ can depend on a number of other factors such as test collection and evaluation metric. For a more detailed discussion of parameter selection, see the recent work of Sakai [28]. In the remainder of this work, we use $\delta = 0.01$.

The 95% confidence interval for increasing values of N is shown in Figure 3. A variation in confidence intervals is observed as values are sampled from $\mathcal{N}(0, \sigma^2)$ in the simulation. Increasing the sample size causes the confidence intervals to narrow, and fit within $\pm\delta$. This exemplifies the encouraged behaviour of expanding the sample to find "statistically significant equivalence". Next methodologies for comparing non-deterministic IR systems with a two

dimensional variance for better effectiveness is discussed.

3. Proposed Methods

In this section, we present methodologies for comparing a non-deterministic system with a deterministic system, as well as with another non-deterministic system. We start with the deterministic — non-deterministic comparison.

3.1. Deterministic — Non-deterministic Comparisons

We propose two novel significance tests to compare a non-deterministic IR system with a deterministic IR system next. One solution is an extension of the standard bootstrap test, and the other is based on multivariate linear modelling.

3.1.1. Bootstrap Test

Algorithm 2 presents the bootstrap based approach for comparing a non-deterministic IR system with a deterministic IR system. The principle is the same in spirit as in Algorithm 1, except that the additional complexity of comparing over M IR system instances of the non-deterministic IR system must be included. Effectiveness scores on each IR system instance are paired by topic with the effectiveness scores of the deterministic IR system. As the effectiveness scores for the same set of topics are now measured on many IR system instances, the t-statistic for the original effectiveness score deltas $t(z)$ is computed from the mean score delta for each topic across M IR system instances. The BS bootstrap observations are drawn from the topic score deltas observed for each IR system instance, and recentered by deducting the mean from the BS bootstrap observations. Evidence in favour of the null hypothesis is found each time the absolute value of the t-statistic for a recentered bootstrap resample $t(z_i^*)$ is greater than or equal to the absolute value of $t(z)$. The final p-value is the count of evidence in favour of null hypothesis divided by $(BS \times M)$.

3.1.2. Multivariate Linear Model Test

The second approach for comparing a non-deterministic IR system with a deterministic IR system uses the following LME model:

$$y_{\ell mn} = \gamma + s_\ell + i_m + t_n + st_{\ell n} + \varepsilon_{\ell mn}. \tag{9}$$

Here $y_{\ell mn}$ is the effectiveness score observed for the $n{\cdot}th$ topic on the $m{\cdot}th$ IR system instance produced with the $\ell{\cdot}th$ non-deterministic IR system, and γ is the model intercept. The factors s_ℓ, i_m, and t_n are the effects due to the $\ell{\cdot}th$ IR system, the $m{\cdot}th$ IR system instance, and the $n{\cdot}th$ topic respectively. The system-topic interaction effect is $st_{\ell n}$. The unallocated portion of effectiveness $y_{\ell mn}$ is what resides in $\varepsilon_{\ell mn}$. The IR system effect (s_ℓ) is fixed in the above model where topics (t_n), IR system instances (i_m) and IR system-topic interaction ($st_{\ell n}$) provide the non-deterministic effects.

The data for the above model consists of effectiveness (y), and three other factor variables for IR system (s), IR system instance (i) and topic (t). Two factors are *crossed* when each level of one factor occurs in every level of another factor, and *nested* if the levels of one factor occurring within the levels of another factor differ. Crossed factors produce an interactive effect in the presence of repeated observations. The effectiveness scores for the same set of topics is measured on each non-deterministic IR system instance and deterministic IR system. Hence, a crossed design can be used where each level of the IR system instance factor for the deterministic IR system is a replicate. In this model, the IR system and IR system instance factors are crossed with the factor topics which result in a system-topic interaction effect ($st_{\ell m}$), but not a system instance-topic interaction effect (it_{mn}) as repeated observations are not available. The p-value for the test is obtained by using the t-statistic for the IR system factor. The degrees of freedom for the test are the number of observations less one.

3.2. Non-deterministic — Non-deterministic Comparison

Now we extend the multivariate linear model test to compare two non-deterministic IR systems. The equation for the model remains the same. The comparison is based on M IR system instances sampled from each non-deterministic IR system. These system instances are different from each other. Therefore, the IR system instance factor (i) naturally nests within the IR system factor (s), as opposed to the crossed design used when comparing against a deterministic IR system. Similar to before the IR system factor is crossed with topics, and causes a IR system-topic interaction effect. The IR system instance-topic interaction effect (it_{mn}) does not occur as the factors are not crossed and have no

ALGORITHM 2: Bootstrap algorithm for comparing a non-deterministic IR system with a deterministic IR system.

$BS \leftarrow$ Number of bootstrap samples;
$N \leftarrow$ Number of topics;
$M \leftarrow$ Number of IR system instances;
$a_n^m \leftarrow$ Matrix of scores for the $m{\cdot}th$ IR system instance of the non-deterministic IR system a on the $n{\cdot}th$ topic, where $m = 1, ..., M$ and $n = 1, ..., N$;
$b_n \leftarrow$ Vector of scores of deterministic IR system b on $n{\cdot}th$ topic, where $n = 1, ..., N$;

// Compute $t(z)$ using mean effectiveness for each topic.
$z \leftarrow []$;
for $n = 1$ *to* N **do**
    $mean_{a_n} \leftarrow \frac{\sum_{m=1}^{M} a_n^m}{M}$;
    $z[n] \leftarrow mean_{a_n} - b_n$;
end
$t(z) \leftarrow \bar{z} / \sqrt{\hat{\sigma}_z^2 / N}$;

// Compute p-value.
$count \leftarrow 0$;
for $m = 1$ *to* M **do**
    $z \leftarrow [a_1^m - b_1, ..., a_N^m - b_N]$;
    $bootstrap\_samples \leftarrow []$;
    $total\_shift \leftarrow 0$;
    for $i = 1$ *to* BS **do**
        $z_i^* \leftarrow$ Resample N items with replacement from z;
        $bootstrap\_samples.append(z_i^*)$;
        $total\_shift \leftarrow total\_shift + mean(z_i^*)$;
    end
    $shift \leftarrow total\_shift / BS$;
    for $i = 1$ *to* BS **do**
        $z_i^* \leftarrow bootstrap\_samples[i]$;
        for $n = 1$ *to* N **do**
            $z_i^*[n] \leftarrow z_i^*[n] - shift$;
        end
        $t(z_i^*) \leftarrow \overline{z_i^*} / \sqrt{\hat{\sigma}_{z_i^*}^2 / N}$;
    end
end
$p\_value \leftarrow count / (BS \times M)$;

repeated observations. The p-value for the test is computed in the same fashion.

4. Synthetic Evaluation of Methods

We now compare the two approaches directly to determine if the two proposed methods for a deterministic — non-deterministic IR system comparison agree. In order to do the comparison, we extend the simulations used in the previous sections. For the deterministic IR system, we randomly pick an effectiveness score between 0 and 1 from a uniform distribution for each topic. For the non-deterministic IR system, we sample a topic effect between 0 and 1 from a uniform distribution, a system instance effect between 0 and 1 from a normal distribution $\mathcal{N}(\mu, \sigma^2)$. We randomly generate the mean μ and the variance σ^2 between 0 and 1 for the non-deterministic IR system. The effectiveness for a topic-system instance pair is the normalized Euclidean distance between the two effects. In order to generate the effectiveness scores for many system instances for a non-deterministic IR system, the topic effect is

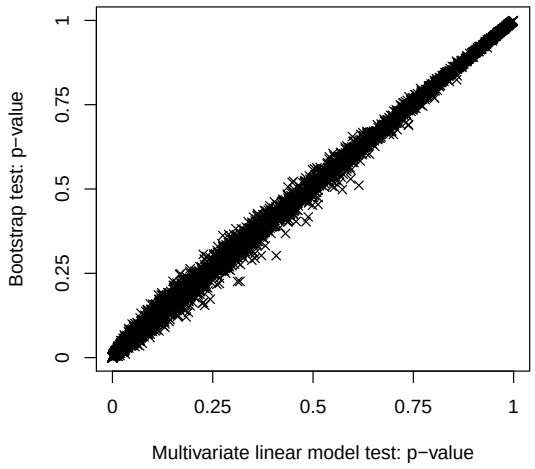


Figure 4: Correlation between bootstrap test and multivariate linear model test when comparing 5000 simulated instance groups with simulated deterministic systems.

fixed, and the system instance effect is sampled repeatedly. We generate effectiveness scores for a set of 50 topics with 100 IR system instances for the non-deterministic IR system, and compared with the effectiveness scores for the deterministic IR system using the two evaluation methods proposed. Figure 4 illustrates the agreement between the two approaches for 5000 such comparisons. Both methods yield similar results.

5. Case Study

We now apply the proposed evaluation methods in a real-world IR scenario. *Sharding* and *tiering* are well-known distributed IR techniques [5]. As the systems tend to be large and complex, a great deal of work has been done to improve the efficiency and the effectiveness in the distributed systems. For example, efficiency can be improved in the system if the query is only sent to a subset of the indexes. However, there is a risk that relevant documents in unsearched shards will be missed. The question becomes how many and which shards should be queried without causing a measurable loss in retrieval effectiveness?

In order to efficiently and effectively select the best subset of shards to visit for each query, the ReDDE algorithm can be used [31]. ReDDE takes a sample of documents from each shard, and uses an index over the sampled shards to select the ordering of shards to search. There are several other resource allocation techniques available but covering all of them is beyond the scope of this paper. See for example Aly et al. [2], Callan et al. [8], Kulkarni et al. [21], Xu and Callan [38]. Here we focus on resource allocation schemes which depend on sampling, such as ReDDE and Rank-S. We refer to any sample-based resource allocation scheme as a *central sample index* (CSI) in the remainder of this paper. The most important take away is that the CSI is a representation of the collection statistics but can vary depending on the samples used.

The simplest form of sharding splits the document collection between shards randomly. Since shards can be searched in parallel, efficiency and throughput can be increased through a variety of mirroring or tiering methods. However, all shards must be searched in order to guarantee the same effectiveness as exhaustive search over the entire collection. This is analogous to a deterministic search in our scenario since the entire collection is searched.

Recent work has shown that the search costs can be reduced by reordering the documents for each shard by topic or similarity [20, 39]. *Topical partitioning* gathers similar documents into a single shard. Now, a search over a carefully selected subset of shards can achieve early precision (P@D, and NDCG@D) closer to exhaustive search without searching the entire collection. The k-means algorithm is one approach to form topically partitioned shards [20, 39]. Each document is represented in a sparse vector space as a *bag-of-words*, where each unique term is a

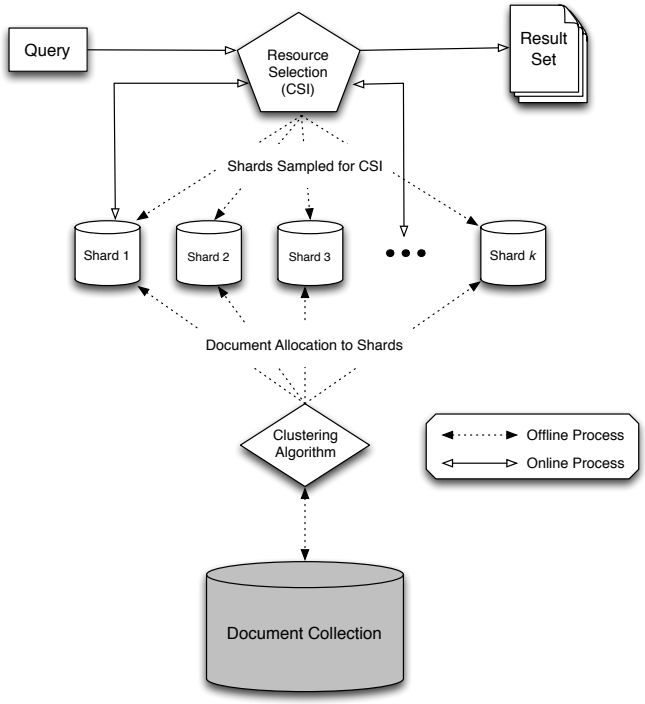


Figure 5: A typical federated selective search system. Shards are created using a clustering algorithms such as *k*-means. In order to create the shards, documents are randomly sampled from the collection and clustered. The remaining documents are then projected onto the *k* shards. Once the shards are formed, representative documents are sampled and indexed in the CSI. Queries are then processed online using a resource selection algorithm such as ReDDE.

dimension. The *k*-means algorithm is not able to scale to large IR collections and therefore sampling is required. Randomly selected seeds from the sampled documents form the initial cluster centroids. Documents are assigned to the closet cluster centroid, based on a *similarity metric*. After all documents are assigned, the center of the assigned documents for each cluster form the new cluster centroids. New random seeds from the sample of documents are used to replace cluster centroids without documents. The process is repeated for several iterations, each time improving the quality of clusters. Finally, all of the remaining documents in the collection are assigned to the closest cluster centroid to form *k* topically partitioned shards. Each time the algorithm is ran, the shard compositions can change dramatically. Figure 5 shows the system composition of a topical partitioning distributed selective search system.

For the purpose of this study, the problem of determining the optimal CSI sample rate is reexamined and follows the process originally presented by Si and Callan [31]. To select a subset of shards for a given query, the CSI is searched first. The proportion of documents from each shard in the CSI search results are used to rank shards for a given query. Therefore, the time spent on searching the CSI is a key factor determining query response time and the CSI search time is correlated to the sample rate used to construct the CSI as well as the query difficulty. The sample rate used for constructing the CSI must be sufficiently high to represent the shard in order to avoid poor retrieval effectiveness. A high sample rate can also minimize the likelihood of encountering out-of-vocabulary (OOV) terms in the mapping of the CSI to the shards. This is a classic effectiveness and efficiency trade-off problem, whereby the best query response time is achieved when the sample rate is set to the smallest level that still achieves similar effectiveness to that resulting from exhaustive search.

5.1. Experimental Testbed

We perform experiments using the TREC GOV2 dataset with TREC topics 701–850 and the ClueWeb '09B dataset

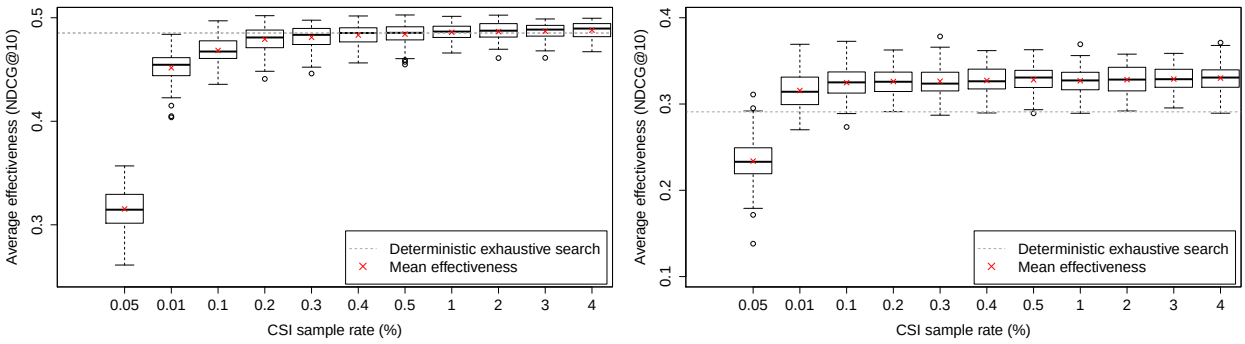


Figure 6: Variation in mean system instance effectiveness observed for the topical sharding algorithm on the TREC GOV2 dataset with TREC topics 701–850 (left) and on the ClueWeb '09B dataset with TREC '09 topics 1–50 (right). Each box in the box-and-whisker plot represents 25'th (Q_1), 50'th (Q_2) and 75'th (Q_3) percentiles of the average system instance effectiveness, while the whiskers span $1.5 \times (Q_3 - Q_1)$ from the box. Any point outside interquartile range is considered an outlier.

using TREC '09 topics 1–50. We use MAP, NDCG@10, and P@10 to evaluate experiments[2]. When not explicitly specified, NDCG@10 is used.

On two independent 1% samples of the document collection, we form 5 topical partitions for each sample using the topical partitioning algorithm proposed by Kulkarni and Callan [20]. Here, a 1% sample is used following prior research by Kulkarni and Callan [20]. Thus, 10 different topical partitions are created and used. As with the original experiments [20], 50 shards per instance for the TREC GOV2 dataset and 100 shards per instance for the ClueWeb '09B dataset are formed, and the full dependency model (FDM) is used to rank the queries [24]. For topically partitioned shards, searching up to 5 shards produced early precision results equivalent to exhaustive search [20]. Therefore, on topically partitioned shards up to 5 ranked shards are searched for each query, except for MAP where a maximum of 10 and 15 shards are searched respectively for TREC GOV2 and ClueWeb '09B datasets. For each of the sharded version, 10 CSI instances are formed for each sample rate, giving 100 instances in total for each sample rate. ReDDE algorithm is used to query shards.

5.2. Results

We first investigate the necessity for a two dimensional significance test. The effectiveness varies for each instances of a non-deterministic IR system, and therefore evaluation for a single instance can be inaccurate. This is illustrated in Figure 6. The figure shows the variation in mean effectiveness for the 100 IR system instances constructed with the topical partitioning scheme for each CSI sample rate on the two datasets. Ignoring variation due to system instances by using a single instance for evaluation is therefore likely to lead to inconsistent conclusions if the sample rate is too aggressive.

So, what is the impact of using conventional significance tests to evaluate non-deterministic IR systems? We compare each individual IR system instance for the topical partitioning scheme with the exhaustive search using a paired t-test, and illustrate the results in Figure 7. As can be seen on both datasets, the number of significant differences with a mean deterioration compared to exhaustive search increases as the CSI sample rates are reduced. Similarly, the number of significant differences with a mean improvement compared to exhaustive search also increases along with increasing CSI sample rates. On both datasets some individual comparisons show a significant difference while others agree. For example, on the TREC GOV2 dataset using a CSI sample rate of 0.5%, 3% of the comparisons show a mean deterioration and a significant difference with a p-value less than 0.05, and 8% with a p-value less than 0.1. For the same CSI sample rate, 2% show a mean improvement and significant difference over deterministic exhaustive search with a p-value less than 0.1, and 90% show no significant difference. This exemplifies the difficulty in evaluating

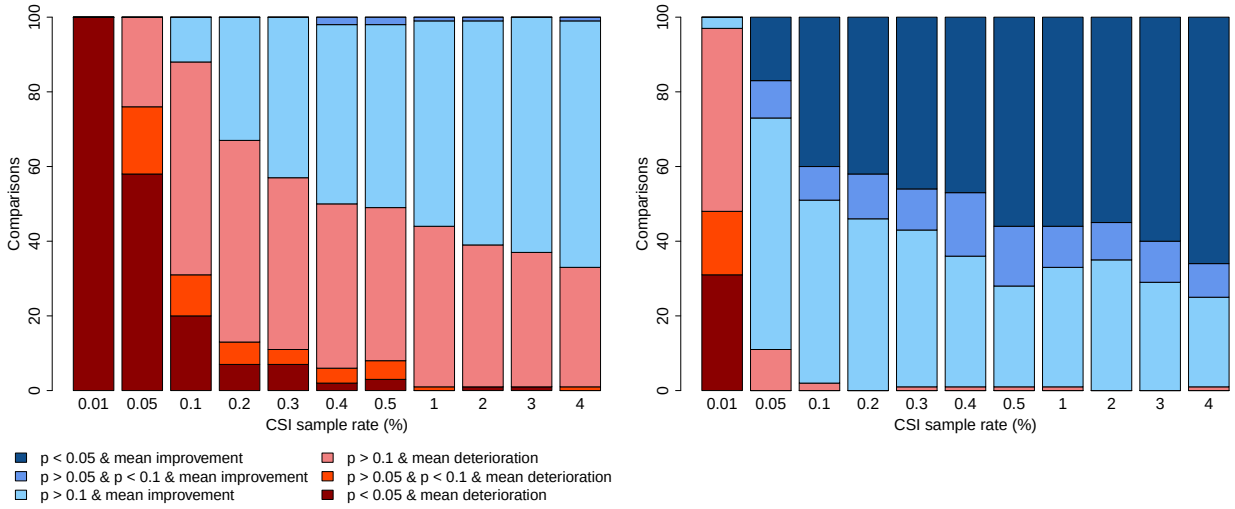


Figure 7: The distribution of p-values for multiple paired t-tests, where each significance test compares effectiveness of an IR system instance derived using the sampling based topical partitioning algorithm for distributed IR system instance setup with exhaustive search on the TREC GOV2 dataset with TREC topics 701–850 (left) and on the ClueWeb '09B dataset with TREC '09 topics 1–50 (right).

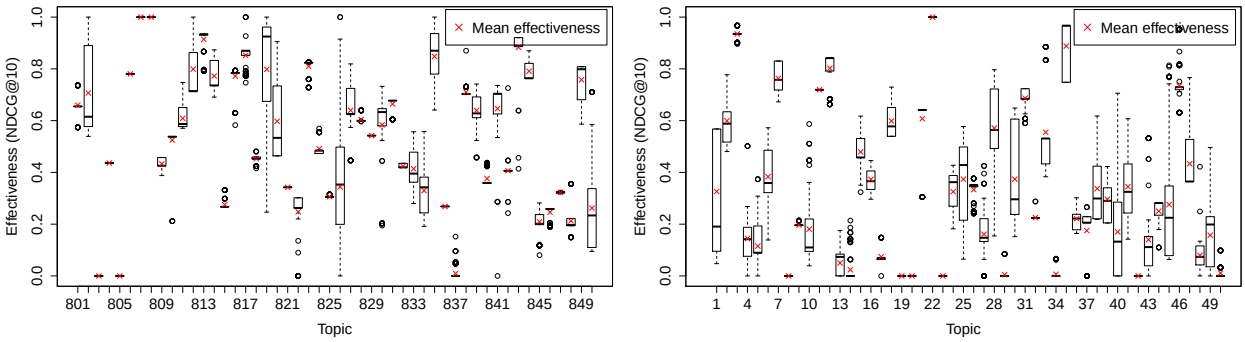


Figure 8: Topical variance for TREC topics 801 – 850 on the TREC GOV2 dataset (left) and TREC '09 topics 1 – 50 on the ClueWeb '09B dataset (right) observed with the 100 IR system instances produced using the topical sharding algorithm with a CSI sample rate of 4%.

systems which use approximation and/or randomization to improve efficiency.

The mean effectiveness for each topic across a large number of non-deterministic IR system instances can be reported to reduce the likelihood of producing inconsistent results. However, such a comparison ignores the variance due to non-deterministic IR system instance effect, and is analogous to comparing two deterministic IR systems just by considering the average effectiveness over a sampled set of topics rather than performing a hypothesis test which takes topical variance into account. We illustrate the variance in effectiveness for TREC topics 801 – 850 on the TREC GOV2 dataset and TREC '09 topics 1 – 50 on the ClueWeb '09B dataset across 100 topically partitioned distributed IR system instances for topical partitioning scheme in Figure 8. While effectiveness for some topics is consistent, others are not. For example, on the TREC GOV2 dataset the TREC topic 826 "Florida Seminole Indians" demonstrate a high variation in effectiveness across system instances, while the topic 836 "illegal immigrant wages" in the same environment displays consistent effectiveness. Similarly, on the ClueWeb '09B dataset TREC '09 topic 1 "obama family tree" shows a high variation in effectiveness, while TREC '09 topic 11 "gmat prep classes" shows no

[2] trec_eval from http://trec.nist.gov is used with default settings to compute the effectiveness of retrieval results.

variation in effectiveness.

There is no obvious reason for these discrepancies, but a few possible reasons can be rationalized. One is the distribution of potentially relevant documents in the CSI. The effectiveness is normally high when potentially relevant documents are explicitly represented in the CSI. Similarly, the effectiveness tends to be low when the CSI sampling does not include these documents. Another possibility is the split of relevant documents across shards. The effectiveness is maximized when relevant documents are concentrated in few shards, and low when they are split between many shards. However, the approaches outlined in this work can help alleviate the problem regardless of the cause.

5.3. Deterministic — Non-deterministic Comparisons

Measuring the effectiveness of a non-deterministic search algorithm using a single instance can never be recommended due to the variance in p-values. The variation in p-value can somewhat be offset by increasing the number of IR system instances in the sample. We illustrate this scenario for the topical partitioning scheme on the TREC GOV2 dataset with TREC topics 701–850 and ClueWeb '09B dataset with TREC '09 topics 1–50 in Figure 9. We generate 50 samples of IR system instance pools for each varying pool size on each dataset using sampling with replacement (bootstrapping). For the experiment, we form 100 IR system instances for each CSI sample rate. These 100 IR system instances for each CSI sample rate are assumed to produce the ground truth for the population. We compare each sample with exhaustive search using the proposed methods for comparing a non-deterministic IR system with a deterministic one. As the graphs for varying CSI sample rates illustrate, having more IR system instances leads to more accurate evaluation.

Recall that the probability of rejecting the null hypothesis when in fact the alternative hypothesis is true is "the power" of the significance test. Plotting the p-values for the number of tests in descending order for each comparison has been a standard practise to illustrate the potential power of significance tests [26, 25]. The p-values for all comparisons performed in the previous experiment on the TREC GOV2 and ClueWeb '09B datasets are arranged in the above manner in Figure 10. Results for each evaluation metric are shown separately, as each graph illustrates a unique set of comparisons. To construct the plot for the standard t-test and the standard bootstrap test, we randomly select an IR system instance from the pool of instances for each comparison. Although the multivariate linear model test and the bootstrap test derive similar p-values as shown in Figure 4, they are not identical. The power of the multivariate linear model test is higher than the bootstrap test for the examples considered here. The difference in the two tests are attributed to fewer assumptions being made about the sampling distribution in a bootstrap test. Both approaches demonstrate a much higher power than either of the standard tests. This indicates the standard tests lack discriminative power due to less observations (data) seen by standard tests than in the two-dimensional tests. Further, more comparisons are found to be significantly different with NDCG@10 than with P@10 for the ClueWeb '09B dataset. This is due to the prominence given to higher ranked relevant documents and inclusion of graded relevance in NDCG@10. A similar observation was made previously by Sakai [27].

Whenever a new evaluation methodology is proposed, validating the approach should be done carefully. In this section, we extended two standard hypothesis testing methodologies from first principles to compare a non-deterministic IR system with a deterministic system. The two approaches use different techniques to compute the p-value. The agreement between the two approaches on the comparisons conducted for previous experiments on TREC GOV2 dataset, and ClueWeb '09B dataset with each evaluation metric are shown in Figure 11. Here, the p-values for comparisons on TREC GOV2 dataset when evaluated with MAP is notable. The topical sharding algorithm shows a significant deterioration with MAP for all comparisons on TREC GOV2 dataset and therefore are located around zero. As shown in the figure, both methods yield consistently correlated results.

5.4. Comparing for Equivalent or Greater Effectiveness

When a new sample-based IR system is introduced for efficiency purposes, a common question is: What is the minimum sample rate required to achieve an equivalent or greater effectiveness than a deterministic baseline? Recall the method used to test for statistically significant equivalence in Section 2. We now extend the method to compare effectiveness of non-deterministic IR systems for similarity.

A possible approach is to apply a multivariate linear model as previously discussed. The goal is to find a setting that is still able to achieve equivalent or greater effectiveness than the exhaustive solution while increasing the efficiency. Hence, only a minimally consequential degradation is considered, so the null hypothesis is $\mathcal{M}_{\mathcal{A}} - \mathcal{M}_{\mathcal{B}} \leq \delta$, instead of

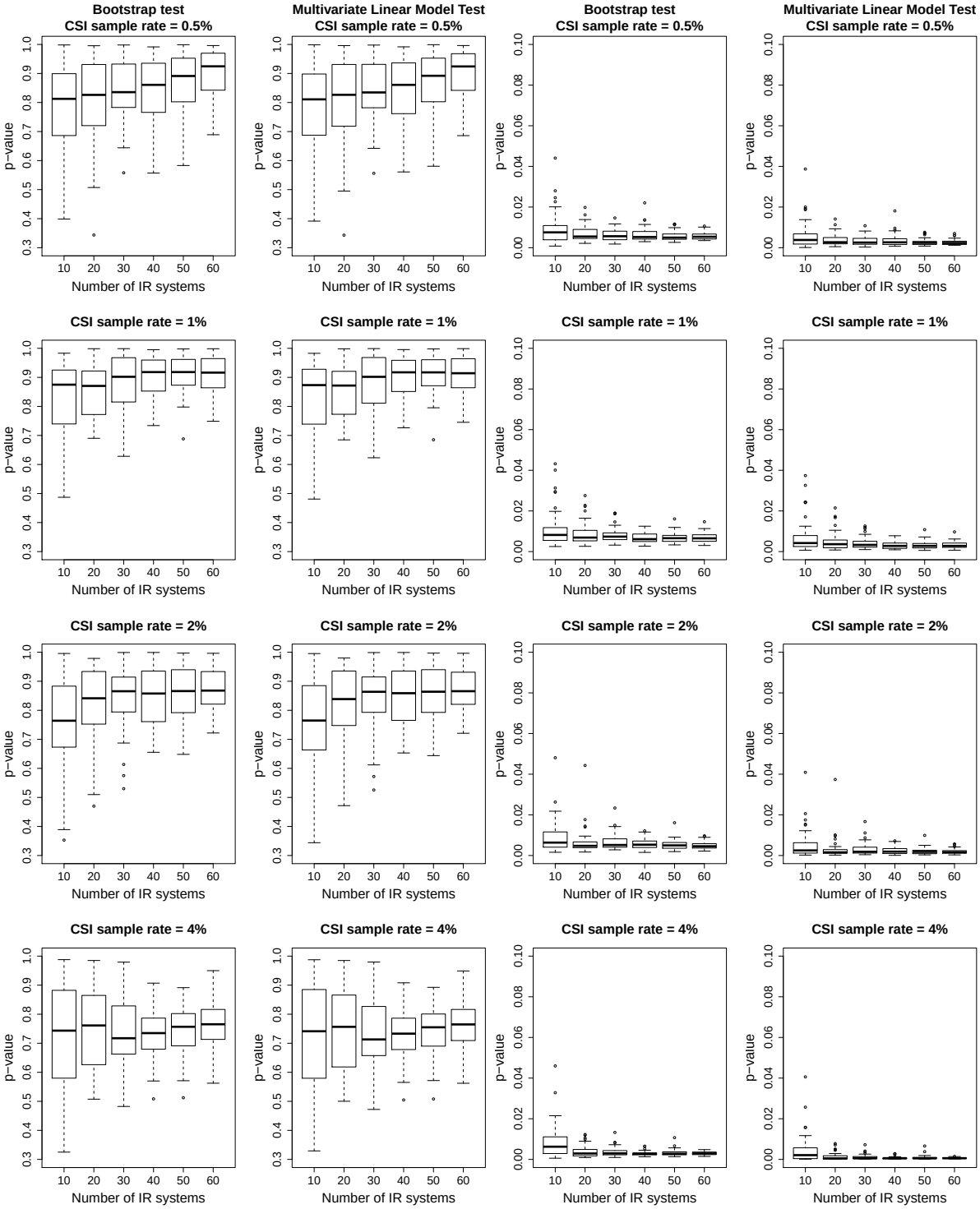


Figure 9: Variation in p-values with proposed tests for varying sample sizes of IR system instances for topical partitioning scheme at different sample rates for CSI with the ReDDE algorithm. The pools of topical partitioning instances are compared with exhaustive search. Experiments are ran on the TREC GOV2 dataset with TREC topics 701–850 (left) and on the ClueWeb '09B dataset with TREC '09 topics 1–50 (right).

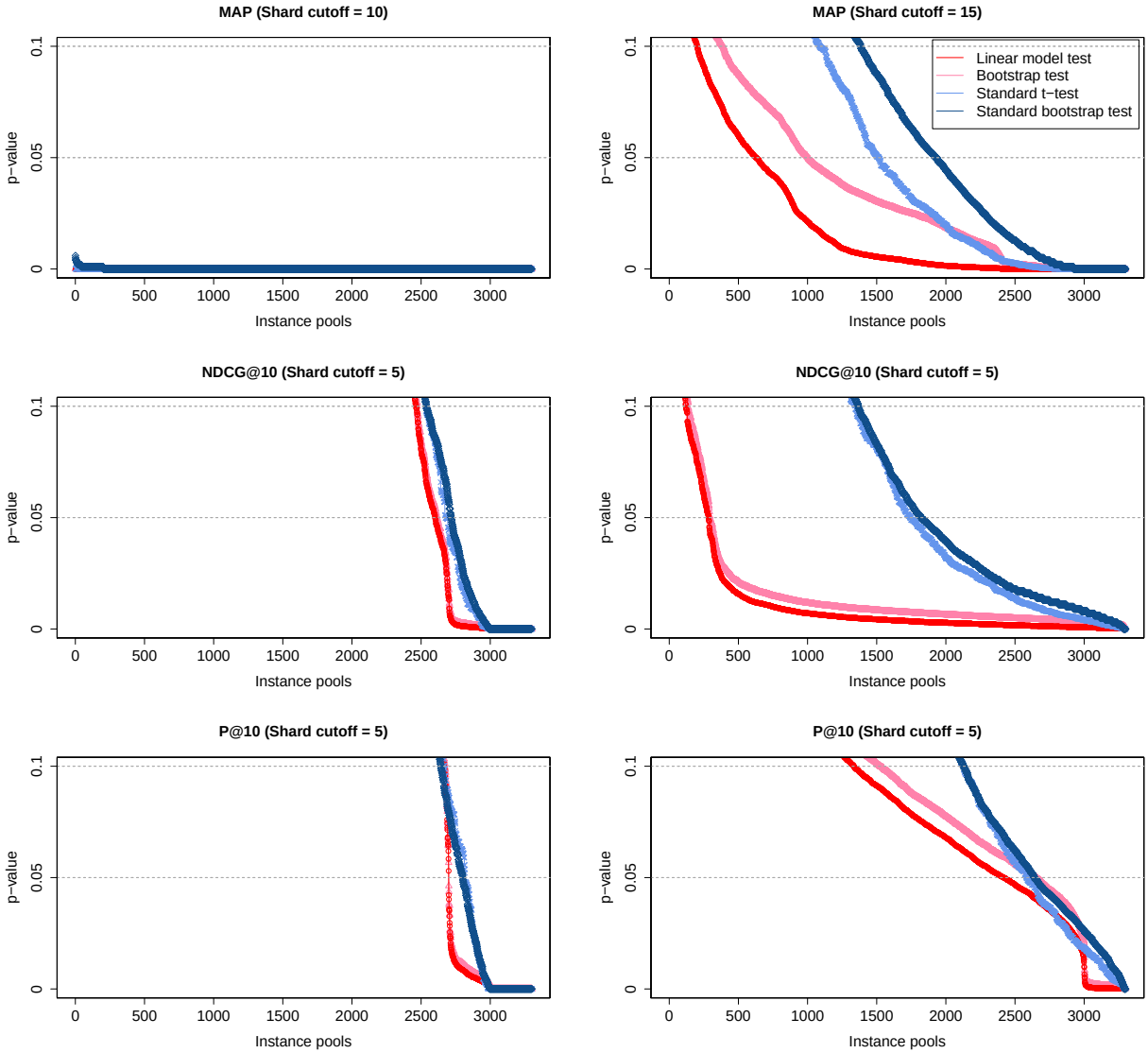


Figure 10: The p-values for different significance tests when comparing 3300 IR system instance pools of topical partitioning scheme with exhaustive search on the TREC GOV2 dataset using TREC topics 701–850 (left) and on the ClueWeb '09B dataset using TREC '09 topics 1–50 (right). Instance pools are sorted in descending order of p-value for each test. For standard significance tests an instance is randomly picked from the pool of instances for each comparison.

$|\mathcal{M}_{\mathcal{A}} - \mathcal{M}_{\mathcal{B}}| \geq \delta$. So $\mathcal{M}_{\mathcal{A}} - \mathcal{M}_{\mathcal{B}} > \delta$ becomes the alternative hypothesis. Here, $\mathcal{A}$ is the non-deterministic approach and $\mathcal{B}$ is the deterministic baseline. For this experiment, a degradation of 0.01 for δ is assumed as to be an acceptable threshold. The null hypothesis is tested, and the smallest setting for which this null hypothesis is rejected is selected. So, effectiveness of a non-deterministic IR system is significantly better than the minimal consequential degradation.

We can indirectly test the null hypothesis using the 95% highest posterior density (*HPD*) confidence interval computed using a posterior distribution for the system factor in the multivariate linear model. If the confidence interval is above δ, the null hypothesis can be rejected, and conclude with a 95% confidence that the effectiveness of the non-deterministic IR system is within δ of the effectiveness of an exhaustive search. However, if δ overlaps with the *HPD* confidence interval, then the null hypothesis cannot be rejected.

We now consider the problem of finding the best cutoff in a selective search system. The 95% HPD interval for the

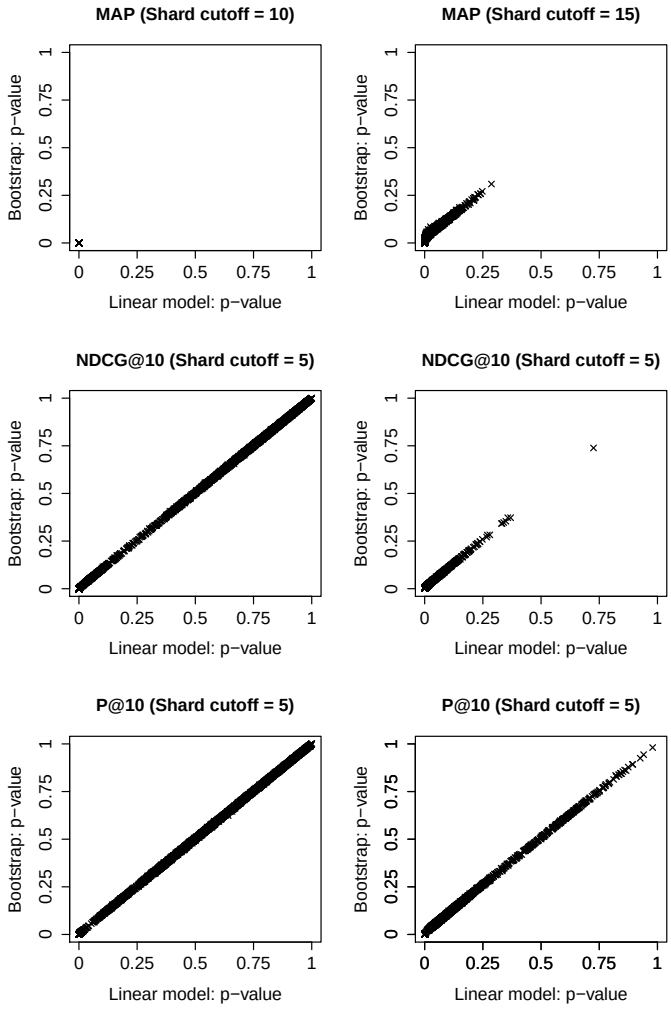


Figure 11: Correlation between a bootstrap test and a multivariate linear model test when comparing 3300 instance pools from the sample based topical sharding algorithm with deterministic exhaustive search on the TREC GOV2 dataset with TREC topics 701–850 (left) and on the ClueWeb '09B dataset with TREC '09 topics 1–50 (right).

scheme as the CSI sample rates are varied for the TREC GOV2 and the ClueWeb '09B datasets is shown in Figure 12. Keeping the focus on NDCG@10, we analyze the results below. The effectiveness is clearly worse than exhaustive search when the sample rate is low. The maximum consequential degradation δ is greater than the confidence interval for sample rates of 0.01% and 0.05% on the TREC GOV2 dataset and 0.01% on the ClueWeb '09B dataset. The δ is within the confidence interval for sample rates 0.1% and above on the TREC GOV2 dataset. So, the non-deterministic approach might not be consequentially worse than exhaustive search, but one cannot conclude this with any certainty. For sample rates $\geq$ 1% for the TREC GOV2 dataset and $\geq$ 0.05% for the ClueWeb '09B dataset, the confidence intervals are above the δ degradation level, and therefore one can conclude that the effectiveness for the non-deterministic method is at least as good or greater than the exhaustive search. The confidence intervals are above zero for a sample rate $\geq$ 0.1% on the ClueWeb '09B dataset. A confidence interval lying above the $randomized - exhaustive = 0$ line implies that the approach performs significantly better than exhaustive search. A pattern similar to NDCG@10 is observed for P@10 on both datasets. Here the effect size can be computed by dividing the difference in mean effectiveness by the standard deviation of the residuals. The effect size at a sample rate of 0.1% on this dataset is 0.52 and 0.41 for NDCG@10 and P@10 respectively. This is an improvement with a medium effect for NDCG@10 and a small

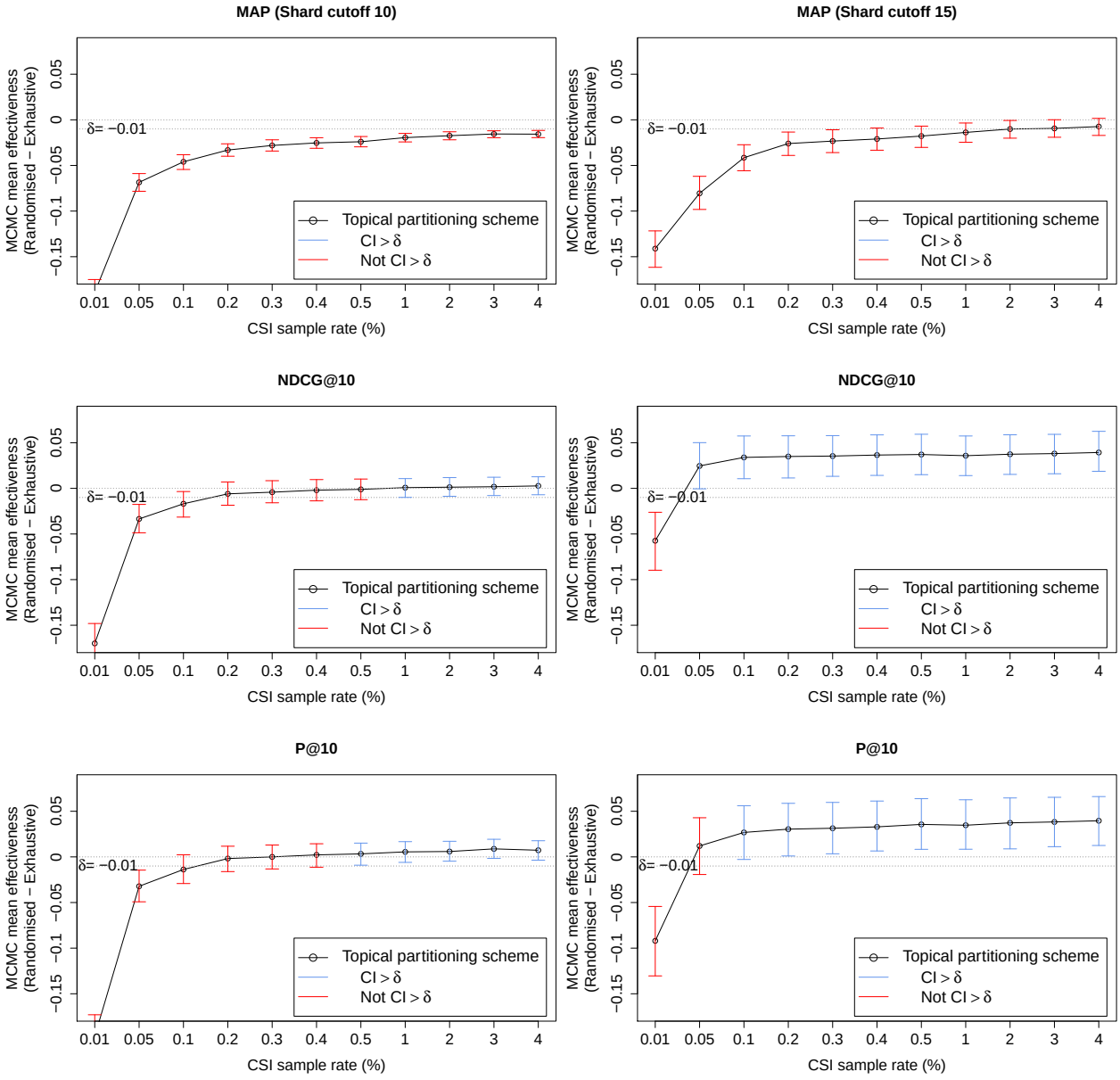


Figure 12: The *MCMC* mean and the 95% *HPD* interval when comparing the non-deterministic IR algorithm with exhaustive search on the TREC GOV2 dataset with TREC topics 701–850 (left) and on the ClueWeb '09B dataset with TREC '09 topics 1–50 (right) with varying CSI sample rates.

effect for P@10. The scheme proposed by Kulkarni and Callan [20] is more effective than the exhaustive approach at a 95% confidence level. Shard cutoffs of 10 and 15 are considered for MAP on the TREC GOV2 and ClueWeb '09B datasets respectively. However, for MAP one cannot conclude with confidence that the effectiveness is no worse than δ for CSI sample rates up to 4%. Here, we illustrated a few applications of the proposed evaluation methods for comparing a non-deterministic IR system with a deterministic IR system. Next, we present results for evaluating two non-deterministic IR systems.

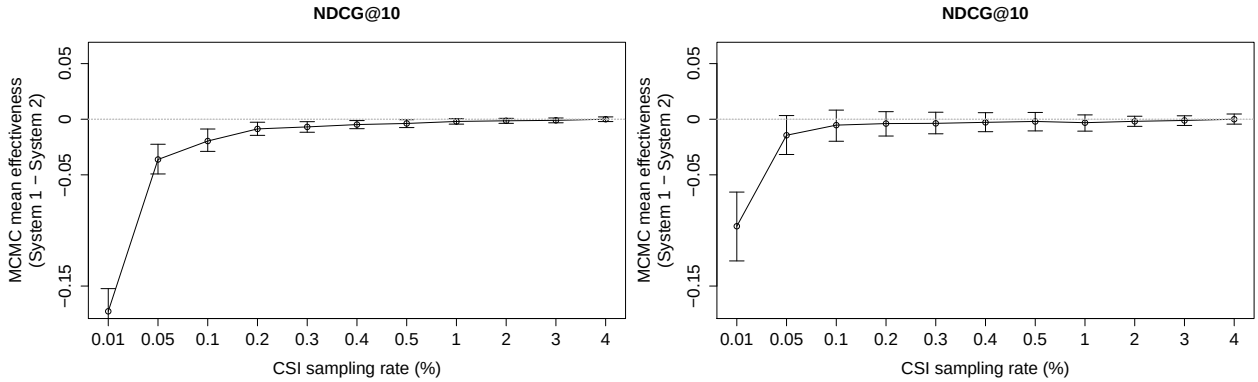


Figure 13: The MCMC mean and the 95% HPD interval when comparing the non-deterministic IR algorithm (System 1) with the same algorithm (System 2). The CSI sample rate is fixed for System 1 at 4%, and varies for System 2. Presented are results for the TREC GOV2 dataset with TREC topics 701–850 (left) and for the ClueWeb '09B dataset with TREC '09 topics 1–50 (right).

5.5. Non-deterministic — Non-deterministic Comparison

We demonstrate the results for separately comparing the topical partitioning scheme at varying CSI sample rates with the same scheme using a CSI sample rate of 4% (another non-deterministic IR system) in Figure 13. The two IR systems are equivalent when the CSI sample rate on the x-axis is 4%. The fact that the comparisons agree as the CSI sample rate approaches 4% for the one with varying CSI sample rate illustrates the validity of the proposed approach for comparing two non-deterministic systems.

6. Conclusion

One important aspect of this work is the recognition of the role multidimensional variance has in evaluation, which occurs when more than one type of experimental unit is being sampled. An example is sampling of system instances and topics for evaluating non-deterministic IR systems. Further, the evaluation can have repeated observations for the same experimental unit, caused by inconsistent experimental conditions. Therefore, variance can be subdivided into many dimensions with the possibility of repeated observations in each dimension.

Despite non-deterministic IR systems with a two dimensional variance becoming increasingly common, little work has been done on how best to compare these systems using traditional evaluation frameworks. We have explored the pitfalls of depending on a single instance of a system. By demonstrating that the effectiveness for the same topic varies across different instances of a non-deterministic system, we motivated the notion of two-dimensional significance testing.

We have proposed two methodologies based on bootstrapping and multivariate linear modelling to compare a non-deterministic system with a deterministic system. Both approaches derive similar p-values, and demonstrate a higher power than current evaluation methodologies. We then extend the multivariate linear modelling approach to compare two non-deterministic systems. Finally, we describe how the proposed methods can be extended to demonstrate that a non-deterministic system has similar performance to another comparable system. The approaches can easily be extended to support repeated measurements for the same topic on the same system instance or for other dimensions of variance.

Future Work: We have explored common scenarios in IR where sampling and approximation can lead to difficulty in evaluation. Similar issues exist in other domains such as machine learning. For example, recommender systems can also have effectiveness vary in two dimensions – users and products. Data analytics is another area where multi-dimensional evaluation could be necessary. We expect to extend this evaluation framework to other areas in future work.

7. Acknowledgements

This work was supported in part by the Australian Research Council (DP130104007). Dr. Culpepper is the recipient of an ARC DECRA Research Fellowship (DE140100275).

References

[1] S. Ahn, S. H. Park, and K. H. Lee. How to demonstrate similarity by using noninferiority and equivalence statistical testing in radiology research. *Radiology*, 267(2):328–338, 2013.

[2] R. Aly, D. Hiemstra, and T. Demeester. Taily: shard selection using the tail of score distributions. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*, pages 673–682, Dublin, Ireland, 2013. ACM.

[3] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan. An introduction to MCMC for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[4] R. H. Baayen, D. J. Davidson, and D. M. Bates. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, 59(4):390–412, 2008.

[5] J. Callan. Distributed information retrieval. *Advances in information retrieval*, pages 127–150, 2002.

[6] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems (TOIS)*, 19(2):97–130, April 2001.

[7] J. Callan, M. Connell, and A. Du. Automatic discovery of language models for text databases. In *Proceedings of the 1999 ACM SIGMOD international conference on Management of data (SIGMOD '99)*, volume 28, pages 479–490, Philadelphia, Pennsylvania, USA, 1999. ACM.

[8] J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of the 18th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '95)*, pages 21–28, Seattle, Washington, USA, July 1995.

[9] B. Carterette. Model-based inference about IR systems. In *Advances in Information Retrieval Theory*, pages 101–112. Springer, 2011.

[10] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM '11)*, pages 611–620. ACM, 2011.

[11] M. Chen and Q. Shao. Monte carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, 8(1):69–92, 1999.

[12] J. Cohen. The earth is round (p ¡ .05). *American psychologist*, 49(12):997, 1994.

[13] J. Cohen. *Statistical power analysis for the behavioral sciences*. Routledge Academic, 2013.

[14] B. Efron. Bootstrap methods: another look at the jackknife. *The annals of Statistics*, pages 1–26, 1979.

[15] B. Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38. SIAM, 1982.

[16] B. Efron and R. Tibshirani. *An introduction to the bootstrap*, volume 57. Chapman & Hall/CRC, 1993.

[17] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web (WWW '01)*, pages 406–414, Hong Kong, 2001. ACM.

[18] R. A. Fisher et al. *The design of experiments.* Edinburgh and London: Oliver & Boyd., 1935.

[19] G. K. Jayasinghe, W. Webber, M. Sanderson, L. S. Dharmasena, and J. S. Culpepper. Evaluating non-deterministic retrieval systems. In *Proceedings of the 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '14)*, pages 911–914, Gold Coast, Queensland, Australia, 2014. ACM.

[20] A. Kulkarni and J. Callan. Document allocation policies for selective searching of distributed indexes. In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*, pages 449–458, Toronto, ON, Canada, 2010. ACM, ACM.

[21] A. Kulkarni, A. S. Tigelaar, D. Hiemstra, and J. Callan. Shard ranking and cutoff estimation for topically partitioned collections. In *Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM '12)*, pages 555–564, Maui, Hawaii, USA, 2012. ACM.

[22] S. Lawrence. Context in web search. *IEEE bulletin of the technical committee on Data Engineering.*, 23(3):25–32, 2000.

[23] F. Liu, C. Yu, and W. Meng. Personalized web search for improving retrieval effectiveness. *IEEE transactions on Knowledge and Data Engineering*, 16(1):28–40, January 2004.

[24] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, pages 472–479, Salvador, Brazil, 2005. ACM.

[25] S. E. Robertson and E. Kanoulas. On per-topic variance in IR evaluation. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '12)*, pages 891–900, Portland, Oregon, USA, 2012. ACM.

[26] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '06)*, pages 525–532. ACM, 2006.

[27] T. Sakai. Alternatives to Bpref. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07)*, pages 71–78. ACM, 2007.

[28] T. Sakai. Designing test collections for comparing many systems. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM '14)*, pages 61–70, 2014.

[29] J. Savoy. Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4):495–512, 1997.

[30] M. Shokouhi. Central-rank-based collection selection in uncooperative distributed information retrieval. In *Advances in Information Retrieval*, pages 160–172. Springer, 2007.

[31] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval (SIGIR '03)*, pages 298–305. ACM, 2003.

[32] L. Si and J. Callan. Unified utility maximization framework for resource selection. In *Proceedings of the 13th ACM international conference on Information and knowledge management (CIKM '04)*, pages 32–41, Washington, D.C., USA, 2004. ACM.

[33] L. Si and J. Callan. Modeling search engine effectiveness for federated search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '05)*, pages 83–90, Salvador, Brazil, 2005. ACM.

[34] P. Thomas and M. Shokouhi. SUSHI: scoring scaled samples for server selection. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR '09)*, pages 419–426, Boston, MA, USA, 2009. ACM.

[35] E. M. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '02)*, pages 316–323, Tampere, Finland, 2002. ACM.

[36] E. Walker and A. S. Nowacki. Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, 26(2):192–196, 2011.

[37] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.

[38] J. Xu and J. Callan. Effective retrieval with distributed collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98)*, pages 112–120, Melbourne, Australia, 1998. ACM.

[39] J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99)*, pages 254–261, Berkeley, California, USA, 1999. ACM.