Evaluating Non-Deterministic Retrieval Systems

Gaya K. Jayasinghe
RMIT University
Melbourne, Australia
gaya.jayasinghe@rmit.edu.au

William Webber
William Webber Consulting
Melbourne, Australia
william@williamwebber.com

Mark Sanderson
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

Lasitha S. Dharmasena
Deakin University
Burwood, Australia
lasitha.dharmasena@deakin.edu.au

J. Shane Culpepper
RMIT University
Melbourne, Australia
shane.culpepper@rmit.edu.au

ABSTRACT

The use of sampling, randomized algorithms, or training based on the unpredictable inputs of users in Information Retrieval often leads to non-deterministic outputs. Evaluating the effectiveness of systems incorporating these methods can be challenging since each run may produce different effectiveness scores. Current IR evaluation techniques do not address this problem. Using the context of distributed information retrieval as a case study for our investigation, we propose a solution based on multivariate linear modeling. We show that the approach provides a consistent and reliable method to compare the effectiveness of non-deterministic IR algorithms, and explain how statistics can safely be used to show that two IR algorithms have equivalent effectiveness.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software— *Performance Evaluation*

General Terms

Experimentation, measurement, information retrieval, effectiveness evaluation, experimental design, statistical analysis

1. INTRODUCTION

How should we measure the effectiveness of an IR system if each run might produce a different output? The obvious solution is to generate several *system instances* and make statistically grounded statements about the overall average effectiveness.

Experiments in IR add another layer of complexity to our problem as retrieval effectiveness varies by topic. The effectiveness of a system 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 such as the t–test or bootstrapping, but these standard tests only support one source of variability (here, in choice of topics). The use of algorithms with non-deterministic output introduces an additional dimension of variability into this scenario. In this paper, we contribute the following

- We present a methodology to solve the two-dimensional significance testing problem (Section 3);
- We explore the properties of our solution on a case study of common sampling-based algorithms – shard construction and centralized resource allocation in distributed IR [3, 6]. We examine the variability that can occur in this environment, observing that an apparently significant result on one instance of a sample-based algorithm can be contradicted by another, and we demonstrate the use of our two-dimensional significance testing methods to handle the variability and provide sound statistical inferences (Section 4).
- We clarify statistical best-practices on parameter selection when comparing algorithms for equivalent effectiveness (Section 5).

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

ACM 978-1-4503-2257-7/14/07 ...$15.00.
http://dx.doi.org/10.1145/2600428.2609472.

2. BACKGROUND

For many years, researchers in the IR community have benefited from shared test collections. A collection includes documents, test topics, and relevance judgments. Using these shared collections, IR systems can be compared by calculating the effectiveness of each using a common metric such as MAP or NDCG. However, arguing that one system is "better" than another 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 set (or population) of possible queries and topics. In response, IR research has adopted the practice of statistical significance testing in order to estimate the likelihood of two systems being hypothetically equivalent (p value). A sufficiently low p value implies a systematic difference that cannot be attributed purely to chance in the selection of topics. Common approaches to significance testing include Student's paired t-test, ANOVA, the sign test, the Wilcoxon signed rank test, Fisher's randomization test, and bootstrapping.

Another approach to testing the significance of differences in system effectiveness on a sample set of topics is through linear modeling. Where variance comes solely through the selection of topics, the following linear model (equivalent to the t-test and ANOVA) is applied:

$$E_{ij} = \gamma + s_i + t_j + \varepsilon_{ij}. \quad (1)$$

Here, E_{ij} is the effectiveness of system i on topic j. We model this effectiveness as due to two factors or effects: the system effect

s_i (essentially, the effectiveness of the system, relative to other systems), and the topic effect t_i (essentially, the difficulty of the topic, relative to other topics). The γ term, the intercept in the model, is the average effectiveness of all systems across all topics, while ε_{ij} is the residual or error term, accounting for the deviation of the observed effectiveness from that explained by the model. The p value for the system effect can either be computed with $\Delta_{s_i}/\sqrt{\sigma^2/n}$ or t-statistic from ANOVA, which follows Student's t distribution with $n-1$ degrees of freedom. Here Δ_{s_i} 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 [2, 5]. For this scenario, sampled topics produce a random effect, and systems produce a fixed effect. Though, not apparent at this stage, 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 the LME with $n-f$ degrees of freedom, where f is the number of fixed effect parameters ($f = 1$ for the above scenario) [1].

Another way to compute the p value is to generate the posterior distribution for the system factor using Markov Chain Monte Carlo (MCMC) simulations. An MCMC simulation repeatedly samples from the conditional distributions of parameter subsets (σ, parameters defining the variance-covariance for random effects, and fixed and random effects) of the linear mixed effect model cyclically, thus making variance of all other parameter subsets reflect the variance for each parameter subset. The posterior distribution of the system effect parameter is expected to follow a normal distribution which can be used to compute the p value [1]. Deriving the posterior distribution also allows us to obtain the highest posterior density (HPD) interval which is analogous to the standard confidence interval. A β % HPD interval represents the shortest interval enclosing $(1-\beta)$ % of the posterior probability mass of the distribution. Therefore, the HPD interval is considered a better representation than the standard error interval.

All of the significance testing approaches above assume an IR system with a deterministic output, where only one observation exists per topic. However, 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 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-system interaction effects as random, the above can be modeled with LME, as follows:

$$E_{ijk} = \gamma + s_i + t_j + ts_{ij} + \varepsilon_{ijk}. \quad (2)$$

Here E_{ijk} is the effectiveness on the k·th observation of system i on topic j, ts_{ij} is the topic-system interaction effect, and ε_{ijk} represents the (random) error of a single observation. However, the model only makes sense if different observations on the same topic-system pair lead to different scores. In Robertson and Kanoulas [5], this variability in topic-system scores is observed over different document sets, whereas in Carterette et al. [2] the variability is in different user types. Our focus in this paper is on non-determinism in IR system output due to variability introduced by, for example, a randomized sampling-based algorithm or an algorithm that exploits logs of (unpredictable) user input. Hence, variability exists in two dimensions (system instances and topics), with one (topical variation) grouped within the other (system instances).

3. OUR APPROACH

Our approach uses the following LME model:

$$E_{lmn} = \gamma + a_l + s_m + t_n + at_{ln} + \varepsilon_{lmn}. \quad (3)$$

Here E_{lmn} is the effectiveness observed for topic n on system instance m generated using algorithm l, and γ represents the model intercept. The factors a_l, s_m, and t_n represent effects for the IR algorithm l, (non-deterministic) system instance m, and topic n respectively. The topic-algorithm interaction effect is captured by at_{ln}. The unallocated portion of effectiveness E_{lmn} is what resides in ε_{lmn}.

The algorithmic effect is fixed in the above model, where sampled topics, system instances and topic-algorithm interaction provide the non-deterministic effects. The data for constructing the above model contains effectiveness (E) and three factors: algorithm (a), system instance (s), and topic (t). If each level of one factor occurs in every level of another factor, we say the two factors are *crossed*. We say they are *nested* if the levels of one factor occurring within the levels of another factor are different.

Crossed factors generally result in an interaction effect in the LME model if they contain repeated measures. The effectiveness for the same set of topics is measured on each system instance and each algorithm. Hence, the algorithm and system instance factors are crossed with the factor topic which results in a topic-algorithm interaction effect, but not a topic-system interaction effect as we do not have repeated measurements.

When two non-deterministic algorithms are compared, the system instances used for evaluation are different for each algorithm, which naturally nests the system instance factor within the algorithm factor. However, if one of the algorithms in the comparison is fully deterministic, a crossed design can be used whereby each level of the system instance factor for the deterministic algorithm is a replicate. The p value for the algorithm factor can be computed with $n-1$ degrees of freedom in the same manner described in Section 2.

Note that our model in Equation 3 differs from the model in Equation 2, as used by Robertson and Kanoulas [5] and Carterette et al. [2]. In the latter model, they regard the observations on the non-topic factor (respectively, document sets, and user type) simply as providing repeat observations of the topic-system interaction, and not as having a systematically grouped effect.

4. CASE STUDY

We now turn our attention to a concrete example. *Sharding* is a well-known technique to divide very large document collections. The technique is used in distributed IR to allocate and index shards of the collection on the different nodes of a cluster. In such a configuration, retrieval is performed across multiple indexes, one per node. Efficiency can be improved if the query is only sent to a subset of the indexes. The question then becomes how many indexes should be queried without causing a measurable loss in retrieval effectiveness. To efficiently and effectively select the best subset for each query, a widely-used approach is to create a centrally held index composed of documents sampled from each shard. This central sample index (CSI) is used to represent the true collection statistics [6].

Recent research has focused on reducing the search cost per query without hurting overall effectiveness by reordering the documents in each shard by topic or similarity [3]. These systems are able to achieve effectiveness close to a search over the entire collection (exhaustive search) while using only a few shards for each

CSI sample rate (%)	t–test p value < 0.05	t–test p value < 0.1	Comparing with exhaustive search (p value)	Comparing two non-deterministic algorithms (p value)
0.01	100%	100%	0.0000	0.0000
0.05	59%	76%	0.0007	0.0000
0.1	21%	31%	0.0601	0.0000
0.5	3%	10%	0.8990	0.0270
1.0	0%	2%	0.9325	0.0991
2.0	1%	2%	0.8748	0.2141
3.0	1%	1%	0.8407	0.3483
4.0	0%	2%	0.7456	1.0000

Table 1: The proportion of system instances that demonstrated a significant difference using a paired t-test, and the p values when comparing the sample-based IR algorithm proposed by Kulkarni and Callan [3] at varying CSI sample rates with a deterministic exhaustive search, and with itself (a non-deterministic algorithm) with a CSI sample rate of 4% using the TREC GOV2 dataset and TREC topics 701 – 850.

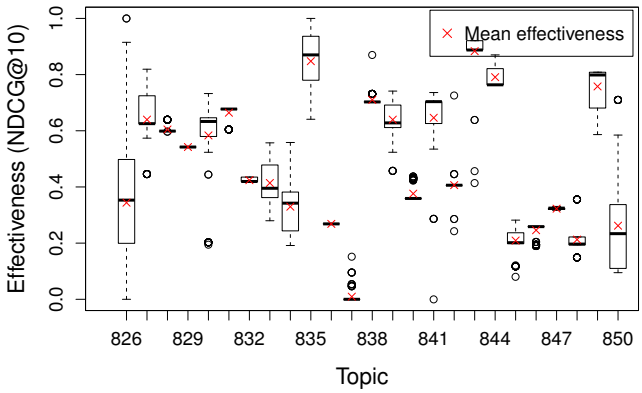

Figure 1: Topical variance for TREC topics 826 – 850 observed with IR system instances produced with the topical sharding algorithm proposed by Kulkarni and Callan [3] on TREC GOV2 dataset with a CSI sampling rate of 4%.

query. However, many clustering and classification algorithms are not able to scale to the typical size of a modern IR collection and therefore resort to sampling and k-means clustering to reorder the shards. Each time such an algorithm is run (i.e. there is a new system instance) distinct shards will be formed.

For the purpose of this study, we reexamine the problem of determining the optimal CSI sampling rate, originally presented by Si and Callan [6]. 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 then used to rank shards for a given query. Therefore, the time spent on searching the CSI is a key factor determining query response time. The CSI search time is correlated to the sampling rate used to construct the CSI and the query difficulty. The sampling rate used for constructing the CSI must be sufficiently high to represent the shard in order to avoid poor retrieval effectiveness. A high sampling 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.

4.1 Experimental testbed

Experiments are performed on the TREC GOV2 dataset, using topics 701–850. Using two independent 1% samples of the dataset, 5 topically-partitioned distributed IR system instances are formed for each sample using a k–means clustering algorithm. Thus, we have 10 different instances of the sharded index. As with the original experiments [3], 50 shards were formed per instance, and the full dependency model (FDM) is used to rank the queries [4]. Selecting a subset of 5 shards produced equivalent retrieval results at depth 10 to exhaustive search [3]. Therefore, only the top 5 ranked shards are searched for each query. For each of the sharded versions, 10 CSI instances are formed for each sampling rate, giving 100 instances in total for each sample rate. NDCG@10 is used as the evaluation metric in all experiments.

4.2 Results

Each individual system instance is compared with the exhaustive search baseline using a paired t–test, and the overall comparison with the proposed approaches are analyzed in Table 1. As can be seen, the number of IR system instances that were significantly different to exhaustive search increases as CSI sampling rates are reduced. Some individual comparisons show a significant difference between two IR system instances while the rest agree on no such difference. For example, at CSI sampling rate of 0.5%, 3% of the comparisons show a significant difference at $\alpha = 0.05$, and 10% at $\alpha = 0.1$. This exemplifies the potential of drawing an inaccurate conclusion, and the difficulty of confidently comparing a single IR system instance produced by a non-deterministic algorithm. We could report the mean effectiveness for each topic across a large number of non-deterministic system instances to reduce the likelihood of producing conflicting results. But this may not result in a fair comparison, as variance due to non-determinism is not explicitly captured in such an evaluation framework.

The variance in effectiveness for TREC topics 826 – 850 across 100 topically partitioned distributed IR system instantiations are illustrated in Figure 1. While effectiveness for some topics are consistent, others are clearly not. A shift in mean effectiveness due to outliers is also observed for several topics. Our proposed approach for significance testing can be used to eliminate such ambiguity and help researchers derive more accurate conclusions.

A comparison of the non-deterministic algorithm at varying CSI sample rates with deterministic exhaustive search and with the same algorithm at a CSI sample rate of 4% is also presented in Table 1. The results verify the suitability of the proposed approach for evaluating non-deterministic algorithms.

We have now discussed the value in using two dimensional statistical significance tests when comparing sample-based algorithms. Assessing similarity when uncertainty is involved even with one source of variability is a concept that has not been broadly addressed by the IR community. Building on the case study, we now outline one possible statistical approach that can be used to compare non-deterministic systems for equivalence.

5. PROMOTING BEST PRACTICE

A statistical significance test cannot be used to "prove" a null hypothesis. The null hypothesis is, in reality, a statistical straw man. Even if we were to believe that two systems have identical population mean scores, we cannot use the methods of inferential statistics directly to perform the test. Statistical inference works probabilistically from the evidence of a sample; exact identity can be determined only by exhaustive examination of the population.

Type	H_0	Desired outcome	Inference	Encouraged behavior
Bad	$M_A = M_B$	Fail to reject	Systems "statistically indistinguishable"	Reduce sample size
Good	$\|M_A - M_B\| \geq \delta_0$	Reject	Difference in systems is significantly less than consequential	Increase sample size

Table 2: Two ways of testing for significance the "equivalence" of system A against system B.

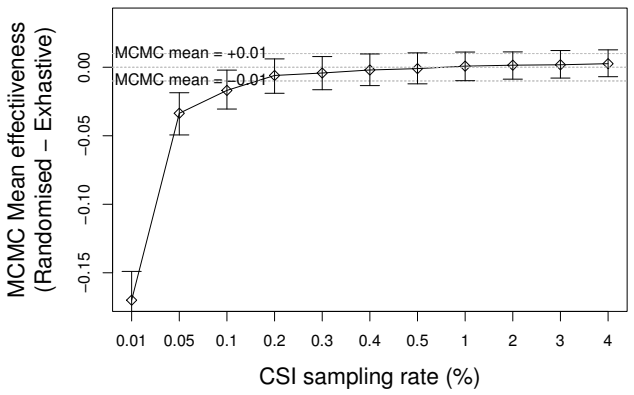

Figure 2: The 95% highest posterior density (HPD) interval for comparison of the topical sharding algorithm proposed by Kulkarni and Callan [3] with exhaustive search on TREC GOV2 dataset and TREC topics 701 – 850.

Even if we believed the null hypothesis, and even if we could establish it statistically, using the failure of a significance test to do so is bad practice. A researcher desires to find no statistical significance (that is, to fail to reject the null hypothesis) in order to confirm an experimental objective. The way to increase the likelihood of this happening is straightforward: decrease the sample size (for example the number of topics or for randomized techniques the number of randomizations), and you decrease the likelihood of finding statistical significance.

If we wish to use hypothesis testing to establish statistically significant "equivalence", a better practice is to proffer the smallest difference in mean performance, δ_0, that we would regard as being consequential. Here, $|M_A - M_B| \geq \delta_0$, becomes the null hypothesis; and $|M_A - M_B| < \delta_0$ is the alternative hypothesis. We then test against the null hypothesis; if the test rejects it, we accept the alternative, and conclude that the difference between system A and system B is no more than δ_0: the systems are effectively equivalent. This process not only makes statistical sense, it also drives good practice: the experimentalist is incentivized to increase the size of the test set and thus the accuracy of the measurements.

The pros and cons of these two methods of testing for statistically significant equivalence are summarized in Table 2.

5.1 Non-deterministic significant equivalence

We now examine our case study. A possible goal is to find the minimum CSI sampling rate that is still able to achieve "equivalent" effectiveness to the exhaustive solution, namely a complete central index. We set $\delta_0 = 0.01$. If a sampling rate with the CSI method causes an absolute difference less than δ_0 in mean NDCG@10 scores, we will regard it as equivalently effective to the full index. Therefore, we test the null hypothesis of $|M_A - M_B| \geq \delta_0$ for each sampling rate, and choose the smallest sampling rate for which this null hypothesis is rejected.

The null hypothesis can also be tested indirectly by computing the highest posterior density (HPD) interval using the posterior distribution for the algorithm factor of the LME model. The HPD interval of the algorithm factor gives a 95% confidence interval on the true difference between the mean performance of the sampled and exhaustive indexes. If the confidence interval is within the lines $\pm\delta_0$ then we reject the null hypothesis and conclude that the system M_B is equivalently effective to system M_A. However, if the confidence interval covers values, below or equal to $-\delta_0$, or above or equal to $+\delta_0$, then the null hypothesis cannot be rejected.

We show the HPD interval for different CSI sample rates in Figure 2. Sampling rates below 0.05% are clearly worse than exhaustive search. For sample rates above 0.1%, a portion of the confidence interval is greater than $-\delta_0$. Therefore, the sampled index *may* not be consequentially worse than the exhaustive one, but we can not draw a conclusion with any confidence. It is not until the sampling rate reaches 1% that the confidence interval is above $-\delta_0$. But, for these sampling rates part of the confidence interval is above $+\delta_0$. Therefore, we cannot conclude with confidence that the sampled method is equivalently effective to exhaustive method. But the fact that the confidence interval is above $-\delta_0$ allows us to say that the method is greater than or equivalently effective (that is, not consequentially worse than) than the exhaustive method for CSI sampling rates of greater than 1%.

6. CONCLUSION

In this paper we have explored the potential pitfalls of depending on a single instance of a non-deterministic system for evaluation. In order to alleviate this problem, we introduce the notion of two-dimensional significance and describe a sound methodology to compare non-deterministic systems. In future work, we will explore how best to safely compare these systems for equivalence directly.

Acknowledgments

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] 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.

[2] B. Carterette, E. Kanoulas, and E. Yilmaz. Simulating simple user behavior for system effectiveness evaluation. In *CIKM*, pages 611–620, 2011.

[3] A. Kulkarni and J. Callan. Document allocation policies for selective searching of distributed indexes. In *CIKM*, pages 449–458, 2010.

[4] D. Metzler and W. B. Croft. A markov random field model for term dependencies. In *SIGIR*, pages 472–479, 2005.

[5] S. E. Robertson and E. Kanoulas. On per-topic variance in IR evaluation. In *SIGIR*, pages 891–900, 2012.

[6] L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *SIGIR*, pages 298–305, 2003.