Validity and Power of t-Test for Comparing MAP and GMAP

Gordon V. Cormack and Thomas R. Lynam
David R. Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
gvcormac@uwaterloo.ca, trlynam@uwaterloo.ca

ABSTRACT

We examine the validity and power of the t-test, Wilcoxon test, and sign test in determining whether or not the difference in performance between two IR systems is significant. Empirical tests conducted on subsets of the TREC 2004 Robust Retrieval collection indicate that the p-values computed by these tests for the difference in mean average precision (MAP) between two systems are very accurate for a wide range of sample sizes and significance estimates. Similarly, these tests have good power, with the t-test proving superior overall. The t-test is also valid for comparing geometric mean average precision (GMAP), exhibiting slightly superior accuracy and slightly inferior power than for MAP comparison.

Categories and Subject Descriptors

H.3.4 [**Information Search and Retrieval**]: Systems and Software – performance evaluation

General Terms

Experimentation, Measurement

Keywords

significance test, validity, statistical power

1. INTRODUCTION

The most commonly reported measure for TREC experiments is Mean Average Precision (MAP), the mean of Average Precision (AP) scores achieved by a particular system for a set of different information needs (topics). The aptness with which MAP characterizes the intended purpose of IR systems is debatable; however, in our experiments we shall assume that MAP (or a similar measure, GMAP [1]) aptly reflects differences in system effectiveness, and consider only the validity and power of statistical tests for the significance of the difference in MAP (or GMAP) between two systems.

Although they are based on questionable assumptions [3], the commonest tests used in comparing MAP are the paired t-test, Wilcoxon signed-rank test, and the sign test. The choice of test is typically based on an uncalibrated tradeoff between validity and power; the assumption being that the t-test is the least valid but the most powerful, the sign test the most valid but the least powerful, and the Wilcoxon test somewhere in between. We present empirical results that fail to demonstrate this tradeoff, and suggest that the t-test is both valid and powerful, and the method of choice.

SIGIR'07, July 23–27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007.

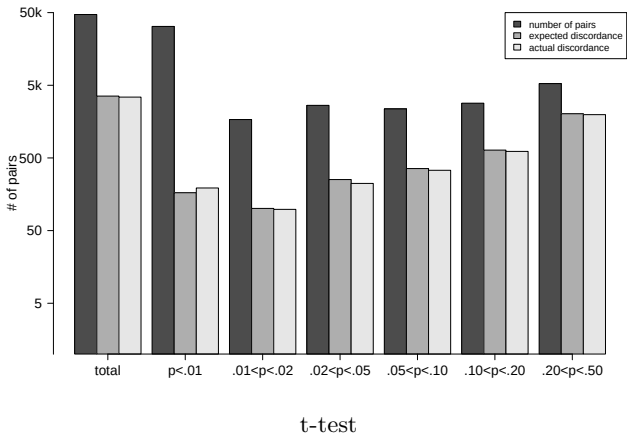

Figure 1: Predicted and actual discordance in MAP comparison ($n = 124$)

2. SIGNIFICANCE TESTING

The goal of measuring inter-system differences is to estimate whether or not system A has a higher *true* MAP than B: "is $M_A > M_B$?[1]" Any particular experiment yields an estimate $\widehat{M_A > M_B}$ of the truth value $M_A > M_B$. Classical statistical tests quantify the significance of this estimate as a p-value. p is the likelihood that a similar experiment might by chance produce the same estimate $\widehat{M_A > M_B}$ even though the opposite ($M_A \leq M_B$) were true. The validity of a statistical test may be characterized by the accuracy with which it estimates this likelihood. However, it is impossible to measure this likelihood directly by experiment, as we can never compute M_A or M_B in order to know the truth value of $M_A > M_B$. It is furthermore difficult to construct a large number of similar experiments, as would be necessary for direct empirical validation.

We therefore validate the statistical tests by comparing the results of pairs of similar but independent experiments,

[1]The true MAP value M_X is defined to be the mean average precision for system X over the universe of all topics.

	pairs	predicted discordance	actual discordance	error	power
$p < .01$	32275 (68.6%)	166	193	+14.0%	0.686
$.01 \leq p < .02$	1688 (3.6%)	101	98	-3.1%	0.721
$.02 \leq p < .05$	2646 (5.6%)	252	223	-13.0%	0.778
$.05 \leq p < .10$	2363 (5.0%)	357	338	-5.6%	~
$.10 \leq p < .20$	2835 (6.0%)	641	617	-4.2%	~
$.20 \leq p < .50$	5273 (11.2%)	2029	1971	-2.9%	~
RMS error				8.6%	

Table 1: t-test Discordance

	rms error	power $\alpha = .01$	power $\alpha = .05$
sign-test	16.4%	0.654	0.757
t-test	8.6%	0.686	0.778
wilcoxon-test	13.8%	0.713	0.801
t-test (GMAP)	8.3%	0.625	0.735

Table 2: Validity vs Power

derived by splitting the set of topics used in a larger experiment – the TREC 2004 robust retrieval track. We use each statistical test to predict d, the probability of a discordant result between the split samples. Note that $d > p$, as d accounts for the sampling error from both splits; p for only one. Over many predictions, the expected number of discordant results is simply the sum of the d values, and, if the test is valid the observed number should be close to this value, invariant when stratified by factors such as the value of p or the magnitude of $M_A - M_B$ (contrary to the apparent results of some previous tests [2]).

It is common practice to deem significant an experimental result with $p < \alpha$ for some fixed threshold α (typically $\alpha = 0.05$). The power of an experimental design is the probability that it will compute a true result with $p < \alpha$. For a valid test, power may be estimated empirically by simulating several experiments and measuring the proportion that yield a correct significant result.

The validity of p should be independent of sample size, magnitude of the difference between the results being compared, and so on. Power, on the other hand, depends directly on both. A larger sample will in general result in lower p-values, and hence increase power. Experimental design must optimize the tradeoff between power and the cost of conducting larger experiments.

3. EXPERIMENTS

The TREC 2004 Robust Track evaluated 110 systems on 249 topics. For each pair of systems we constructed several random equal splits with 124 topics per split, and applied three statistical tests – paired t-test, Wilcoxon signed-rank test, and sign test – to one of the splits. Using the test we computed both p and d. We summed the values of d, stratified by p, and also counted the number of discordant results between the two splits. The results for the t-test are presented in figure 1 and table 1. Of the 47080 t-tests, 32275 (68.6%) yielded $p < .01$. Of these tests, predicted and actual discordances totalled 166 and 193, a difference of 14%. Other strata of p contained fewer tests and resulted in smaller errors. The RMS error over all strata was 8.5%. We take this low value to validate the t-test. Power depends on the chosen α value; for $\alpha = 0.01$, power is 0.68, for $\alpha = .05$, power is 0.778.

Table 2 shows RMS error and power for four statistical tests: t-test (repeated from table 1), Wilcoxon test and signed test applied to difference in MAP, as well as t-test applied to difference in GMAP. We note that the sign test has higher error and lower power than the t-test while the Wilcoxon test has higher error and marginally higher power. t-test applied to GMAP shows a comparable error rate to t-test applied to MAP, and somewhat lower power.

Topics	rms error	power $\alpha = .01$	power $\alpha = .05$
25	14.7%	0.394	0.555
50	13.0%	0.533	0.664
75	7.0%	0.606	0.716
124	8.6%	0.686	0.778
249	-	0.775	0.844

Table 3: t-test Validity vs Power

A second set of experiments used unequal splits to measure the sensitivity of the t-test to the number of topics sampled. For this experiment, we assumed that the t-test for the larger sample was accurate (as evidenced by the first experiment) and combined it with the t-test for the smaller sample to estimate d. Any increased error, therefore, could be attributed to the smaller sample size. Table 3 shows error and power as a function of sample size. As expected, error rates were somewhat higher for smaller sample sizes but overall predicted discordance agrees very well with actual discordance. Power increases with sample size, as expected.

4. REFERENCES

[1] Robertson, S. On GMAP and other transformations. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management* (New York, NY, USA, 2006), ACM Press, pp. 78–83.

[2] Sanderson, M., and Zobel, J. Information retrieval evaluation: Effort, sensitivity, and reliability. In *SIGIR Conference 2005* (Salvador, Brazil, 2005).

[3] Van Rijsbergen, C. J. *Information Retrieval, 2nd edition*. Dept. of Computer Science, University of Glasgow, 1979.