Improvements That Don't Add Up: Ad-Hoc Retrieval Results Since 1998

Timothy G. Armstrong, Alistair Moffat, William Webber, Justin Zobel

Computer Science and Software Engineering
The University of Melbourne
Victoria 3010, Australia

{tgar,alistair,wew,jz}@csse.unimelb.edu.au

ABSTRACT

The existence and use of standard test collections in information retrieval experimentation allows results to be compared between research groups and over time. Such comparisons, however, are rarely made. Most researchers only report results from their own experiments, a practice that allows lack of overall improvement to go unnoticed. In this paper, we analyze results achieved on the TREC Ad-Hoc, Web, Terabyte, and Robust collections as reported in SIGIR (1998–2008) and CIKM (2004–2008). Dozens of individual published experiments report effectiveness improvements, and often claim statistical significance. However, there is little evidence of improvement in ad-hoc retrieval technology over the past decade. Baselines are generally weak, often being below the median original TREC system. And in only a handful of experiments is the score of the best TREC automatic run exceeded. Given this finding, we question the value of achieving even a statistically significant result over a weak baseline. We propose that the community adopt a practice of regular longitudinal comparison to ensure measurable progress, or at least prevent the lack of it from going unnoticed. We describe an online database of retrieval runs that facilitates such a practice.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*.

General Terms

Experimentation, Measurement, Standardization.

Keywords

Retrieval experiment, evaluation, system measurement, survey.

1. INTRODUCTION

Information retrieval (IR) research has a strong tradition of empirical evaluation, stretching back to the Cranfield experiments of the 1960s [Cleverdon, 1967, 1991]. These established the standard methodology for assessment of retrieval effectiveness: a test collection consisting of a fixed document corpus, a set of topics or queries, and judgments indicating which documents are relevant to which topics. To measure a retrieval system, the queries are run against the corpus, returning a ranked list of documents or *run* for each query. The group of runs a system returns for a set of topics will be referred to here as a *runset*. The runs are marked up for relevance using the judgments, and the relevance vectors are scored using a measure such as mean average precision (MAP) or rank-biased precision (RBP) [Harman, 1993, Moffat and Zobel, 2008].

In collaborative experiments, multiple groups submit systems that are compared against each other. But in laboratory work, the typical scenario considered in this paper, as few as two systems may be used, one implementing a new technique, the other providing a baseline for comparison. The validity of the new technique is tested by comparing its score to the baseline, and any apparent improvement is then tested with a statistical significance test [Zobel, 1998, Sanderson and Zobel, 2005, Smucker et al., 2007].

Creating test collections is costly. The document corpus must be collected; topics must be formulated; and, most expensive of all, relevance judgments must be performed. However, once a test collection has been created, using and reusing it is cheap, as no further human involvement is required in the evaluation process. Thus there is a strong incentive for researchers to reuse existing test collections rather than create new ones. While there are objections that could be made to this continuous reuse of the same test data, it does offer one great advantage: it allows us to compare the results of different research experiments, so long as they are performed on the same collection. If different researchers perform experiments on different, privately-formed collections, comparing scores is problematic [Webber et al., 2008]. However, if researchers run experiments against the same collection, direct comparison of scores is straightforward and informative.

The IR research community has available to it several large-scale, high-quality test collections created through collaborative experiments, most notably the collections generated by the TREC effort [Voorhees and Harman, 2005]. Founded at the start of the 1990s, TREC is an annual experiment involving many research groups working on a range of retrieval tasks. The TREC effort collects sizeable document corpuses, formulates topics, and undertakes relevance assessments, creating judgment sets that are tolerably comprehensive even for multi-gigabyte document sets [Zobel, 1998]. The dozens of systems used in test collection construction are a rich resource both for analysis of the collection itself and for comparative evaluation of subsequent retrieval innovations. As a result, the community is in the enviable position of being able to conduct experiments that are deterministic, completely repeatable,

CIKM'09 November 2–6, Hong Kong, China

and produce results that can be directly compared with those generated by other research groups in different times and places. Because of these virtues, the TREC collections are widely used in retrieval experiments, forming a reference point that can be employed to compare different retrieval techniques over time.

In this paper, we report an analysis of published results from retrieval effectiveness experiments run against TREC collections over the past decade. We investigate the reported results, and also examine aspects of experimental practice such as the choice of baseline system against which the new technique is to be compared. In practice, the innovation is usually implemented on top of the baseline system, and together they represent a "with" and "without" pair for some proposed technique. Our analysis covers all results against the Ad-Hoc (including Robust), Terabyte, and Web collections of TREC, published in the ACM SIGIR Annual International Conference on Research and Development in Information Retrieval since 1998, and in the ACM CIKM Annual International Conference on Information and Knowledge Management since 2004.

The key question we sought to investigate was whether, and by how much, IR techniques have improved over the period surveyed, as reflected in the reported effectiveness scores achieved against standard collections. As we have available to us the scores achieved by the systems that participated in the original TREC experiments, we can use these as a benchmark to assess subsequent systems. We can also ask how competitive the baselines are, and whether over time the systems used as baselines incorporate new discoveries and thus represent the state of the art fairly. Finally, we can also ask the methodological question of whether researchers are in fact reporting results in a way that makes them easily comparable; specifically, whether they are indeed using the standard collections, and employing them in the standard way.

The results of our analysis are not encouraging, particularly for the Ad-Hoc and Robust collections. In summary:

- Baselines are rarely competitive, lying in general in the second quartile or below of the original TREC experimental systems. Baselines do not seem to incorporate new discoveries in the field, as they do not improve over time.

- Statistically significant improvements are often claimed, but few of the new systems are competitive with the best of the original TREC results. Indeed many "improvements" fall below the median of the original TREC systems.

- Most worryingly of all, there is no discernible upward trend in Ad-Hoc scores over time. Rather, the pattern is of researchers consistently reporting similar improvements over similar baseline scores, with results reported in 2008 generally indistinguishable from those reported in 1999. Matters are slightly better for the (more recent) Web collections, but even so there is no consistent upwards trend.

These findings are all the more surprising in that the reuse of test collections gives a strong comparative advantage to later systems: the researchers have access to the topics and judgment sets, know all the experimental results achieved previously, and have ample time to tune their systems.

What are we to make of these results? They appear to demonstrate that ad-hoc IR technology has not improved over the past decade, as indeed some commentators have suggested. If so, this lack of improvement needs to be more widely and clearly recognized, especially since few would suggest that ad-hoc retrieval has been perfected. Some might argue that, on the contrary, improvements are meaningful even if demonstrated against weak baselines, especially if significance has been achieved; and that it is up to the system integrator, not the researcher, to aggregate all of these disparate improvements into a single, outperforming system. Such an argument rests on the degree to which improvements are additive, a question which we explore empirically as a contribution of this paper. Even if improvements are independent, the magnitude of the effect they generate may diminish as the baseline quality improves. In particular, all that some particular improvement might be doing is compensating for a defect elsewhere in the mechanism, and, if that other defect is eradicated, then the improvement in question might be rendered impotent, or possibly even counter-productive.

We believe that responsibility lies with researchers to show that some innovation they propose has led to an overall improvement in effectiveness. Indeed, our analysis is open to the disturbing interpretation that we are observing a perverse selection bias, in which researchers who attempt to improve on competitive benchmarks fail to achieve significant improvements, and so do not succeed in publishing their work; while those working off weak baselines achieve statistical significance more easily, and publish more readily. As a community of scholars, we should expect that competitive baselines be used, and that researchers demonstrate the relationship between their work and the strongest prior retrieval techniques, not just the most convenient ones.

Whether ad-hoc IR performance has indeed stopped progressing, or whether there is still improvement being made, but not being adequately demonstrated, it is clear that historical data needs to be both more accessible and more cited. We end this paper by describing a public system that we have deployed to facilitate and validate longitudinal comparisons of retrieval effectiveness against standard test collections, so as to support this goal.

2. RELATED WORK

The aims, methods, and achievements of the TREC effort are described in Voorhees and Harman [2005] and in the proceedings of the conference. Of particular relevance is the overview document from the proceedings of TREC-8 [Voorhees and Harman, 1999], the last TREC at which the Ad-Hoc track was run. The decision to discontinue the track is explained as being due to a plateauing of improvements in ad-hoc retrieval. This plateauing is illustrated by running the versions of the SMART system that participated in each of the eight Ad-Hoc tasks against all the eight test collections created for those tasks. The results showed no improvement in MAP scores for SMART after the version that participated in TREC-5 [Voorhees and Harman, 1999]. These results are only conclusive for the SMART system itself; they do not demonstrate that other TREC systems did not continue to improve. The same figures are reproduced in tabular form in Buckley [2005], with the conclusion that since TREC-5 there have "only been minor improvements in SMART". Similarly, Lynam et al. [2004] remark that ad-hoc retrieval effectiveness had reached a plateau by 1999, but suggest that additional performance gains may be achievable.

In Armstrong, Moffat, Webber, and Zobel [2009b], we compare the normalized effectiveness of systems participating in the TREC-3 to TREC-8 Ad-Hoc Track and the TREC-2003 to TREC-2005 Robust Track, which used a similar test collection and methodology. We run five publicly available retrieval systems in a total of seventeen different configurations against the nine test collections. The scores of these reference systems are then used to standardize (see Webber et al. [2008]) the scores of the original TREC runs, in order to control for variability in test collection difficulty and allow comparison across test collections. We observe no improvement in the retrieval effectiveness of median, first quartile, or best TREC systems between 1994 and 2005, as shown in Figure 1.

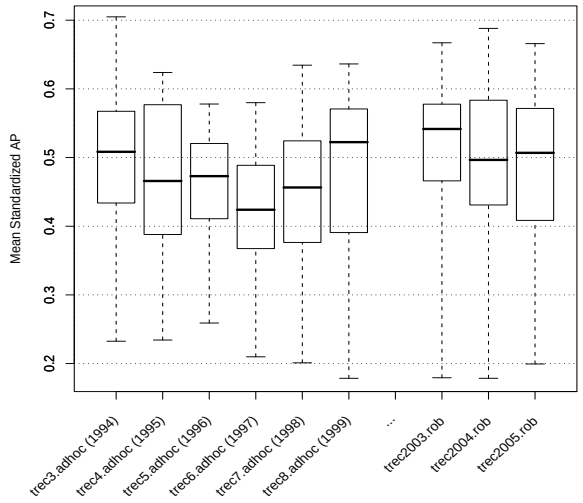


Figure 1: Mean standardized AP scores of runsets submitted to nine TREC events, excluding manual systems, with standardization factors established by a pool of 5 current public systems and 12 of their variants, plus 2 background systems. The central line in the box is the median score; the top and bottom of the boxes are the quartiles. This figure is reproduced from out poster paper presentation, Armstrong et al. [2009b].

Practice in the literature regarding citation of previous results is variable; we describe some illustrative examples. Liu et al. [2005] note that their method achieves the best published scores for the TREC-2004 Robust test collection on both new and all topics, and cite the previous best scores; however, they do not note that they also achieve the best published scores for the TREC-6 and TREC-7 collections, nor do they note their standing on the TREC-8 collection, where they come fifth, though they are the best title-only run. At the other end of the performance spectrum, de Vries and Roelleke [2005] use some of the weakest baselines in our analysis (a MAP of 0.138 for a title-only run on TREC-7, and of 0.183 on TREC-8). The baseline system is derived from a re-implementation of an earlier method [Hiemstra et al., 2004], which the authors cite. The authors also note that the baseline system performs worse than their system achieved at the TREC-8 competition, but do not note the relative positions the baseline scores would have achieved at TREC (bottom 10% of automatic title-only runsets at TREC-7, bottom 15% at TREC-8). Finally, to take an example with mid-range performance, Fang and Zhai [2006] implement their own baselines over which they achieve statistically significant improvements. The baselines are close to the median of the original TREC systems, but no TREC scores or other previous results on the same collections are given.

3. SURVEY OF PUBLICATIONS

We have analysed retrieval results reported on TREC ad-hoc style collections. All SIGIR papers in the period 1998–2008, and all CIKM papers in the period 2004–2008, were examined. SIGIR is the premier venue for the presentation of innovations in IR, so this is where we expect to find the results most indicative of the overall state of IR research. In recent years the CIKM conference has become a significant destination for publications in the field, so the last five years of CIKM were also included.

Results were recorded for conference papers that presented effectiveness scores for ad-hoc style retrieval on TREC collections, meaning the TREC Ad-Hoc, Robust, Web, and Terabyte collections, and subsets or combinations thereof. We also included cases where judgments from the TREC routing track were incorporated to enable the use of a larger subcollection, for instance, for WSJ or AP subcollections. We collected results for MAP and precision at depth 10 (P@10), the two most commonly reported evaluation metrics, and the only ones sufficiently common in publications to permit meaningful longitudinal analysis. Papers were excluded if:

- Parameter optimization had been performed, without a separate held-out test collection used for evaluation.
- Topic relevance judgments were used as part of query processing (as is the case in true relevance feedback).
- There was reason to believe that the reported results were inaccurate (such as incompatible MAP and P@10 scores).
- The test collection used was not sufficiently documented.

The following information was captured for each paper that was included in the survey:

- MAP and P@10 scores listed for each test collection for the new method being presented. We denote these as the *improved* scores, to distinguish them from baseline scores. Improved scores were generally reported in tables of results, but in cases where it was possible to determine the score from a graph, it was also recorded. Where multiple sets of scores were provided, the following rules were applied:
  - results reported for different topic lengths (for example, title only or description only) were captured separately;
  - where scores for several system variants were reported, scores for the best variant were captured;
  - where ambiguity remained, we captured the result that appeared to best reflect the performance of the method presented.
- MAP and P@10 *baseline* scores corresponding to the above, including any referenced or cited scores relating to a "good" technique that were used in the paper as a basis for a claim of improvement (whether the paper explicitly described it as

SIGIR 1998	112, 206, 275
SIGIR 1999	90, 191, 214, 222, 246, 254, 309
SIGIR 2000	10, 345
SIGIR 2001	1, 35, 111, 120, 181, 334, 390, 414
SIGIR 2002	3, 49, 283, 417, 425
SIGIR 2003	4, 159
SIGIR 2004	64, 138, 178, 186, 194, 266, 440, 448, 482, 486, 540, 552, 564
SIGIR 2005	19, 226, 242, 250, 282, 298, 465, 480, 605, 661
SIGIR 2006	75, 91, 115, 139, 154, 162, 178, 621
SIGIR 2007	175, 271, 295, 303, 311, 319, 383, 391, 599, 679, 729, 759, 777, 843
SIGIR 2008	67, 171, 227, 235, 243, 419, 427, 443, 491, 817, 821, 825, 855
CIKM 2004	32, 42
CIKM 2005	305, 307, 321, 331, 525, 672, 688, 704
CIKM 2006	550, 559, 800, 866
CIKM 2007	253, 545, 711
CIKM 2008	399, 1417, 1431, 1441

Table 1: Conference proceedings and starting page numbers of papers with results included in our survey.

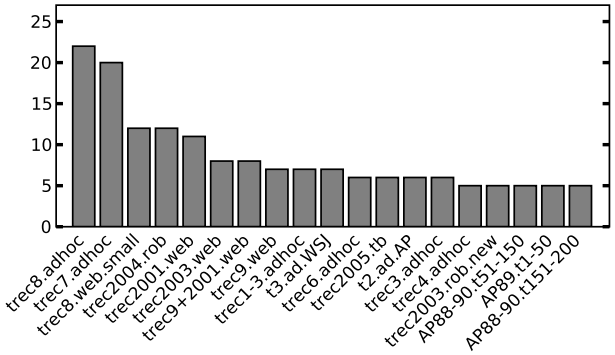


Figure 2: Frequency of usage of TREC test collections, including variant collections, in SIGIR and CIKM papers. Usages are counted as the number of publications reporting at least one score for that exact collection. Only collections used in 5 or more publications are shown. Another 101, or 38% of usages are not shown, as the collection appeared in less than 5 publications.

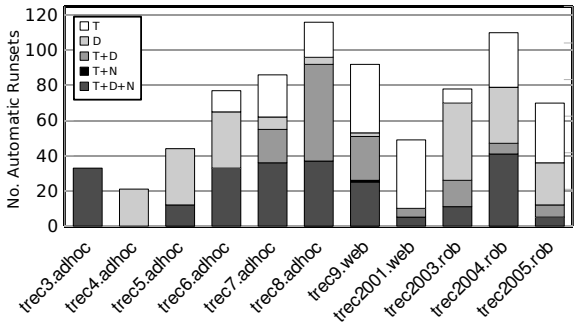


Figure 3: Breakdown of submitted runsets by the topic fields used (title, T; description, D; and narrative, N) for query generation for selected TREC tracks in the years 1994 to 2005.

a baseline or not). In cases where multiple baselines were present, the strongest was selected.

- Whatever (if any) "best known" MAP or P@10 scores cited in the paper for that collection.
- The nature of the predominant claim being made in the paper, categorized as being one of *effectiveness* (better quality results); *efficiency* (less resources required to compute the same results); *distribution* (better load sharing or better scalability); or *none* (covering cases where the results did not exceed the baseline, or the paper was focused on comparing existing methods, rather than presenting new ones).
- Whether there was a claim made about the statistical significance of improvements over (any of, if there were multiple) baseline results.

A total of 85 SIGIR papers and 21 CIKM papers met the analysis criteria; they are listed in Table 1. Of these 106 papers, 89 were focused on retrieval effectiveness, 7 on efficiency, 5 on distribution, and 5 on other issues. Results from all these types of research focus are included in the following analysis. The set of papers studied includes four that had authors in common with this paper.

One surprising result of the analysis was the number of variant test collections used, despite our restriction to the fairly homogeneous tasks and collections of the TREC Ad-Hoc, Robust, Web, and Terabyte tracks, and despite the desirability of using common experimental collections to aid comparability. Figure 2 shows the most frequently used collections. In the 106 publications that were surveyed, a total of 83 different test collections had been used, 40 of them only once. These 83 collections were created from 30 different combinations of document corpuses and 35 different combinations of topic subsets. Of the 83 test collections, 70 were derived from the same five-disk Tipster corpus used extensively in several TREC tracks including the Ad-Hoc and Robust tracks. Many of these Tipster-derived collections were closely related. For example, 15 variations on the Associated Press subcollection and 10 variations on the Wall Street Journal subcollection were present.

What proportion of runsets used which topic fields to generate queries varied from year to year between different TRECs, as shown in Figure 3. For example, the focus of the TREC-2001 Web Track was on short title-only queries, and the majority of submitted runsets accorded with this expectation. The length of query used in subsequent experimentation is thus a possible confound to our longitudinal study, and was a factor noted for each reported result encountered in the SIGIR and CIKM papers.

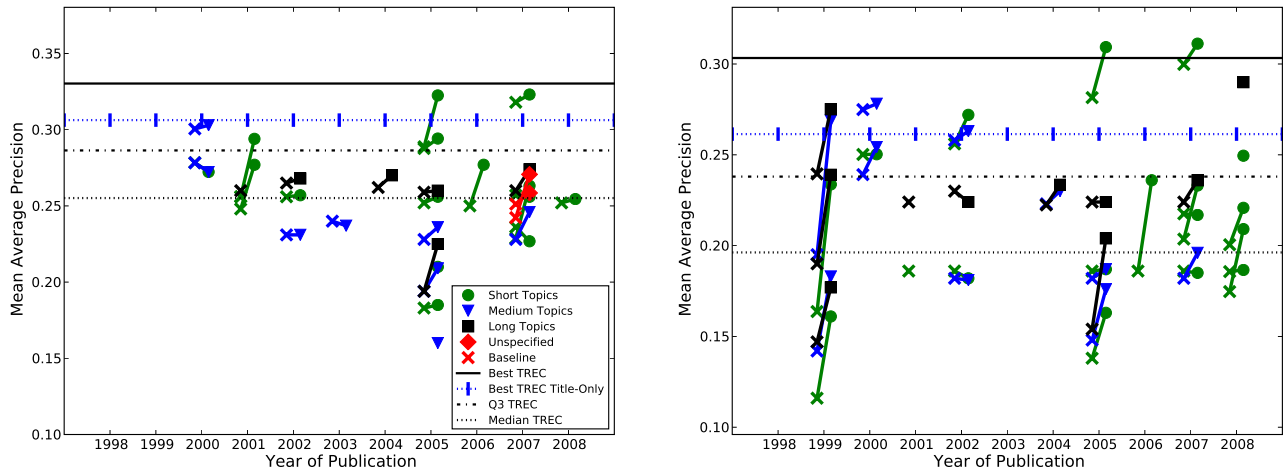
4. ANALYSIS OF REPORTED RESULTS

Results for four of the most frequently used test collections are presented in Figure 4; eight less frequent collections are given as an appendix in Figure 8. We would ideally hope to see that the original TREC runs provide a foundation for steady upward gains in both baseline and improved outcomes. Achievements in one year should translate into better baselines one or two years later, with the "best original run" being eclipsed by new techniques in a process of continual improvement, in the way that Olympic swimming records may stand for a while, but are always eventually bettered.

The most widely used collection is TREC-8 Ad-Hoc, cited by 22 papers, and with 30 result pairs. The reported results are displayed in Figure 4(a), and distinctly differ from the ideal structure hypothesized above. Immediately obvious is that the baseline scores are generally uncompetitive. Over half the baseline scores for TREC-8 are below the median score achieved by the original automatic TREC systems in 1999, a situation that is still continuing nearly ten years later. Only four baselines from the nine-year survey period are in the top quartile, and only one baseline is even close to the best of the original TREC submissions. But the latter must surely be regarded as being the natural starting point for subsequent work, particularly so when the original runs have been available online via TREC throughout the decade. Nor is there any upward trend in baseline scores over time. The mean baseline score prior to 2005 is 0.260; from 2005 onwards it is 0.245.

A similar lack of progress can also be observed in Figure 4(a) for the improved scores on the TREC-8 Ad-Hoc collection. Certainly, improved scores are mostly better than the corresponding baselines. But the improved scores do not trend upwards over time; and only five of the 30 improved scores are in the top quartile of the original 1999 automatic TREC systems. Over the whole decade only two title-only systems beat the best automatic TREC title-only system, and no system beats the best automatic TREC system across all query types. The apparent conclusion is that this decade of published papers, and the experiments they report, has not resulted in improved retrieval effectiveness.

The results for the TREC-7 Ad-Hoc collection are shown in Figure 4(b) (36 result pairs from 20 papers), and tell a similar story. Neither baselines nor improved scores trend up over time. Indeed, the mean of each score type is lower from 2005 onwards than prior to 2005. Most baselines are below the median score



(a) TREC-8 Ad-Hoc ($n = 32$)

(b) TREC-7 Ad-Hoc ($n = 38$)

(c) TREC-8 Small Web ($n = 21$)

(d) TREC-2003 Web (Topic Distillation) ($n = 8$)

Figure 4: Published MAP scores for four different TREC environments, as reported in papers in the SIGIR and CIKM Proceedings. The connections show comparable before-after pairs, that is, the baseline (offset to the left) and improved scores reported in a published paper. Best overall and title-only, upper quartile, and median automatic runsets are marked. The guideline for the best original TREC-8 Small Web title-only run is a lower bound, since not all submissions indicate the topic part used.

of the original 1998 TREC automatic systems, and most of the improved scores are inferior to the original 1998 system that delimited the top quartile. Only three title-only systems beat the best automatic TREC title-only system, and only two systems beat the best overall automatic TREC system. Results for the TREC-3 and TREC-6 Ad-Hoc collections, and for the similar TREC-2003, TREC-2004, and TREC-2005 Robust collections can be found in Figure 8. They show the same typical pattern: weak baselines, few improved scores that outperform the best TREC systems, and little indication of an upward trend over time.

The only scores that exceed the best automatic TREC system are the TREC-6, TREC-7, TREC-2004 Robust, and TREC-2005 Robust results reported in Liu et al. [2005] and Zhang et al. [2007], and a TREC-4 score reported in Mitra et al. [1998].

Figure 4(c) shows results for the TREC-8 Small Web collection; results for the TREC-9 and TREC-2001 web collections, individually and combined, can be found in Figure 8. The results on these ad-hoc style web collections differ from those on the Ad-Hoc proper and Robust collections mainly in having stronger baselines, with almost all baseline scores being above the median for the original TREC runs, and several being in the top quartile. There are also more reported systems that outperform the best original automatic TREC system. There is, however, still no long-term upward trend. The combination of initial improvement with subsequent stagnation suggests that the most fertile development period for a retrieval problem may be shortly after that problem is proposed.

The only collection with MAP results that regularly better the original TREC submissions and demonstrate an upward trend in performance is that of the Topic Distillation task of the TREC-2003 Web track, with seven results pairs from seven papers, shown in Figure 4(d). Here, six improved scores beat the best TREC system, as do two baselines. There also appears to be an upward trend in performance of both baselines and improved scores over time, albeit on a small sample. This trend may be due to the relative newness of the Topic Distillation task, in its second year at TREC-2003. Again, newer tasks may inspire greater progress.

Results for P@10 are in all cases similar to those seen for MAP, and are not shown – partly for space reasons, partly because P@10 was reported less frequently than MAP, and partly because P@10 is in general regarded as being a less reliable metric than MAP. Less

frequently cited collections are also not reported, as they contribute little information about trends over time.

5. RUNNING ON THE SPOT?

An external view of how research on ad-hoc retrieval should have proceeded over the past decade would look for a process of verifiable, cumulative improvement in technology over time. The use of the standard TREC collections would allow experimental results to be compared without re-experimentation being required. Each research group would build on previous discoveries, with one year's best system being the next year's baseline, and all of them building from the foundations laid by the original TREC participants. As the technology matured, progress might be increasingly incremental. And there might be questions about whether the stereotyping of methodology made such incremental progress meaningful. Nevertheless, there would be measurable improvement.

Our analysis, however, demonstrates that this picture of verifiable progress is not what has occurred during the last ten years of ad-hoc retrieval research. Baselines are inconsistent, generally uncompetitive with the original TREC systems, and do not become more demanding over time. Improved systems are rarely competitive with TREC's benchmark, and show no upwards trend. This apparent lack of improvement is consistent with our earlier finding that systems participating in the TREC Ad-Hoc tracks have not become stronger at least since TREC-3, in 1994 [Armstrong et al., 2009b]. There is, in short, no evidence that ad-hoc retrieval technology has improved during the past decade or more. Each year, researchers report statistically significant results; but each year, the baselines that significance has been achieved against are the same.

There are several possible explanations for the lack of an upward trend in performance. One is that there simply has been no improvement in ad-hoc retrieval technology for more than a decade, but that researchers, reviewers, and readers have been unaware of this because of faults with experimental methodology – and thus have submitted and accepted failed attempts to improve ad-hoc IR. (And, worryingly, as a community we may have favored publication of papers where authors made use of a poor baseline, because these papers are the ones that appear to show the most dramatic improvements.) The underlying issue may be that ad-hoc retrieval has reached a plateau, at least using current approaches. In this case, the urgent task becomes to correct the systematic faults that have obscured the lack of progress. In Section 7, we propose changes to methodology that address these faults.

An alternative explanation might be that real improvements in technology are being made, and that this is demonstrated by the fact that most individual experiments on effectiveness-oriented methods show an improvement over a baseline, and more than a third of them claim significance. If we take this approach, then, provided the improvements are original, it should be sufficient to demonstrate them over a simple baseline, rather than having to go to the expense and complexity of creating a state of the art system. What we care about, this view states, is demonstrating that there is an improvement, not achieving optimal performance. It is up to the system developer, not the researcher, to integrate all these improvements into a single, high-performing system. Such a viewpoint must be discomforted by the lack of a trend in improvement over time, in both research and TREC results – surely if real improvements were being made, at least some of them would eventually rub off on successive experimental systems. More concretely, this line of argument (which is not one that we agree with) raises the question of whether techniques that demonstrate improvement in isolation are additive in combination. We make an approach to this issue in Section 6.

A third explanation would be that researchers are in fact aware that the new methods being presented do not provide overall improvements in retrieval effectiveness, and have proposed them for other reasons – because, for instance, they are more theoretically elegant, or require less information to be stored or processed, or are more efficient in some other way, or demonstrate some other interesting behavior. The main purpose of subjecting these methods to an experimental evaluation of retrieval effectiveness (apart from meeting the expectations of reviewers) is to demonstrate that they achieve comparable effectiveness to sane baselines; that is, that the proposed improvements do not significantly harm retrieval effectiveness. Since such research is published, we can presume it is worthwhile. Even here, though, the use of less than competitive baselines is open to question. After all, a change that doesn't harm a weak baseline may nevertheless harm a stronger system. We investigate this issue in the following section.

6. ADDITIVITY OF IMPROVEMENTS

A question posed in the previous section is whether improvements over weak baselines are meaningful, even where they are statistically significant. How confident are we that a technique that yields an improvement over a weak baseline would also give an improvement over a strong one, and therefore be a worthwhile addition to state of the art systems?

We approach this question through the simple model of a retrieval system as a set of techniques, drawn from a universe of $\mathscr{T} = \{t_1, t_2 \ldots t_n\}$ of n existing techniques. In this model, techniques are arbitrarily combinable, though not necessarily orthogonal in their effect. A system R is defined by which techniques $T_R \subset \mathscr{T}$ it incorporates, with two such systems of particular interest: the vanilla system V which implements no techniques, $T_V = \emptyset$; and the state of the art system S with the combination of existing techniques (not necessarily all of them) that yields the greatest effectiveness, T_S : $\forall T \subset \mathscr{T}, Eff(T_S) \geq Eff(T)$. This model is of course simplistic, but it facilitates ready experimentation. A researcher develops a new technique, t_{n+1}, implements it on top of the vanilla system T_V, and demonstrates experimentally that the new system $T_V \cup \{t_{n+1}\}$ outperforms the vanilla baseline. The question is, how confident can we be that $Eff(T_V \cup \{t_{n+1}\}) > Eff(T_V)$ implies $Eff(T_S \cup \{t_{n+1}\}) > Eff(T_S)$. That is, does an improvement over a vanilla baseline imply an improvement over a state of the art system?

To directly answer this question for the improvements proposed in the literature surveyed would involve implementing not just those improvements, but also systems that could be shown to be state of the art at the time the improvements were proposed – an ambitious undertaking. Our approach instead is to take an existing, publicly available system with a range of options that can be switched on and off, to simulate adding or omitting a technique. By evaluating the performance of every combination of these options, we can explore the question of whether improvements are reliably additive, and whether a technique that enhances a baseline will also enhance a more advanced system. However, we need to be cautious about the conclusions we draw from this experiment. There is a strong selection bias here: the only techniques available to our experiments are those already implemented in the public system, and the very fact that they have been implemented means that they are likely to be exactly those techniques that offer improvement in a wide range of different settings. Thus, this experiment is more likely to overstate than understate the additivity of techniques.

The system chosen for this experiment was Indri, as bundled with version 4.8.0 of the Lemur toolkit. [1] Six options were identi-

[1] `www.lemurproject.org/indri`

Toggle	Enabled	Disabled
Term Smoothing	Dirichlet Prior [Zhai and Lafferty, 2004].	Jelinek-Mercer.
Ordered Phrases	Ordered proximity windows, with a maximum of 4 terms between each occurence, scored for every sequence of 2 or 3 terms in the original query [Metzler and Croft, 2005]. Tuning resulted in a weighting of 0.1/1.0.	No ordered proximity.
Unordered Proximity	Unordered proximity windows, with a maximum size of four times the number of terms being scored, for every sequence of two or three terms in the original query [Metzler and Croft, 2005] (This diverges slightly from the original method. described in the paper, but the number of possible combinations grows exponentially with query length). Tuning resulted in a weighting of 0.1/1.0.	No unordered proximity.
Query Expansion	Pseudo relevance feedback, using Indri's adapted version of relevance modelling [Lavrenko and Croft, 2001] with a total of twenty terms selected from ten documents, weighting the original query as 0.3 and the expanded query 0.7.	No query expansion.
Stemming	Porter Stemming.	No stemming.
Stopping	Stopping using the standard list of 417 stopwords included in Indri.	No stopping.

Table 2: Indri options toggled for additivity experiment. The configuration with all toggles on was very similar to the run *indri05AdmfS* at the TREC-2005 Terabyte track, which achieved the top MAP score for a title-only runset (although the difference with several other top runsets was not statistically significant at $\alpha = 0.05$) [Clarke et al., 2005, Metzler et al., 2005].

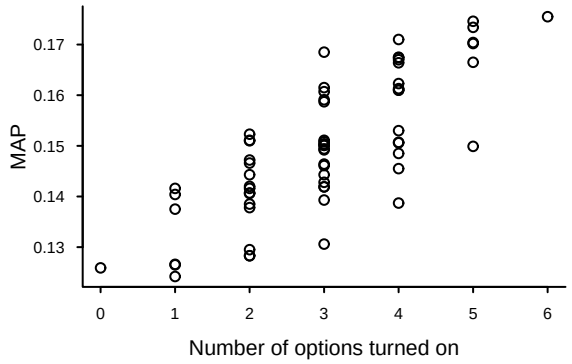


Figure 5: MAP as a function of number of options turned on, for Indri running against the TREC-5 Ad-Hoc test collection.

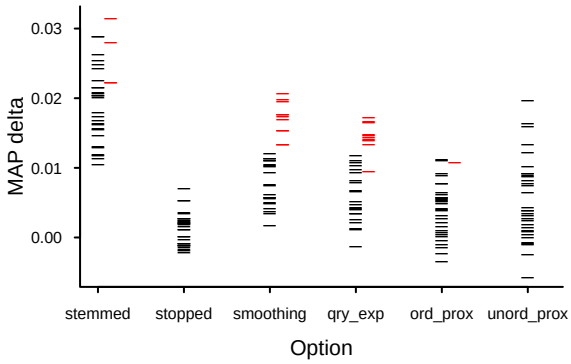


Figure 6: Improvement in mean average precision from turning an option on, for Indri running against the TREC-5 Ad-Hoc test collection. Each point represents the delta in MAP between the specified feature being turned on and being turned off, with the settings of all other features being held the same. Improvements that achieve significance in a paired, two-tailed, two-sample *t* test ($\alpha = 0.05$) are offset to the right. Every combination of features is considered.

fied that offered some improvement in performance. These options are described in Table 2. Under our model, each option represents a technique, with one setting of the option representing the absence of the technique, the other its presence. The test collection used was the TREC-5 Ad-Hoc collection. Each of the $2^6 = 64$ different combinations of techniques (options) was run against the collection, and the resulting runsets scored using MAP. (Experiments were also run with Indri and the same six options against the TREC-2001 Web collection, and with Terrier 2.2 and five options on the TREC-5 Ad-Hoc collection. In both cases, the results were similar overall to those reported here, although which option showed what degree of benefit did vary.)

Figure 5 plots the MAP scores achieved by the different Indri configurations in the TREC-5 Ad-Hoc environment, as a function of the number of options turned on. There is a positive relationship between the number of options turned on and the retrieval effectiveness achieved, suggesting that, here at least, options are broadly additive. Additionally, there is no obvious tendency for adding options to have a weaker effect when more options are set (say, going from four options to five options) than when fewer are set (say, going from one option to two options).

Figure 6 gives a different viewpoint of the same experiment. Here, we show the effect of adding a single technique, with every other combination of options held fixed. Since for each option there are $2^5 = 32$ different combinations of the other five options, for each option we record 32 different MAP deltas resulting from turning that option from "off" to "on". The point to notice is the improvement that an option offers depends upon the combination of other options that are enabled at the time, and is highly variable. There are instances where an option creates a significant improvement when added to certain configurations, but has no overall effect when added to others. So, while improvements are additive on average, they are not additive always, and additivity needs to

be confirmed in individual cases. Note again the caveat – this is with techniques that have been selected for implementation in this system due to their demonstrated value.

It is worth pausing to consider what is meant by "statistical significance" in the context of Figure 6. The significance of a new technique's improvement is routinely established by sampling (actually or in assumption) across topics, presuming that the corpus, and the baseline system to which the technique has been added, are fixed. But Figure 6 raises the question of whether significance also needs to be established across the variety of different possible system configurations, as the strength (and even the sign) of the effect depends upon the configuration that the technique is added to. Of course, calculating significance across configurations is problematic, because different configurations of the same system are not independent of each other. Also, the different configurations of a system vary in importance; showing improvement over the current state of the art setup is more compelling than showing it over a vanilla baseline.

Testing a new technique against a range of configurations helps to establish the generality of its benefit. Many improvements are clearly not additive; two query expansion methods are unlikely to provide further improvement when combined, or a novel length normalization may simply be inapplicable alongside an effective similarity measure such as BM25. On this reasoning, it seems evident that the onus is with the researcher to show that their method can improve systems that are already effective, and that – as is true across science – reviewers have the responsibility for properly scrutinizing those claims. In the case of ad-hoc IR we suspect that both components of this partnership have weakened in recent years.

7. A PUBLIC DATABASE OF RUNS DATA

The critical experimental failing, in our view, is that the great majority of papers only report on experiments that the researchers have carried out themselves, without reference to past results. It is our view that this practice is unacceptable, and has led directly to the issues reported in this paper.

With the widespread use of standard test collections and evaluation metrics, it is straightforward in principle to provide a repository for IR system runs and effectiveness results that would allow comparison between results submitted by different research groups at different times. Such a repository would go a long way to addressing the shortcomings already discussed in this paper, by promoting transparency about reported results. Even if authors eschew the use of prior data, making a database of results readily available to referees would in itself raise community standards.

We see the requirements of such a repository as including:

- Use of the submitted TREC runs to provide a consistent reference scale for "standard" experiments. Indeed, one of the stated purposes for the TREC Ad-Hoc track was that it "documents the state of the art and provides a basis of comparison" [Voorhees and Harman, 1999].
- The ability for researchers to add new test collections, built either as combinations of old resources, or as entirely new resources. For example, we would hope that the same tool would cater for CLEF and NTCIR comparisons.
- Support for a range of effectiveness metrics, including retrospective application of new ones as they are developed.
- Support for a range of statistical tests between systems.
- Support for additions to the set of runs. In particular, when a paper is submitted, the researcher could be expected to commit their runs to the repository on a "private" basis, and be issued in return with a URL that lists attributes of that run and compares it with published effectiveness scores achieved in that test collection. That URL would then be included in the submitted paper, to be used by the referees to examine the run's details and performance.
- Support for permanent "publication" of runs, so that when a submitted paper is accepted and published, the runs data will be permanently available, and accessible by other users.

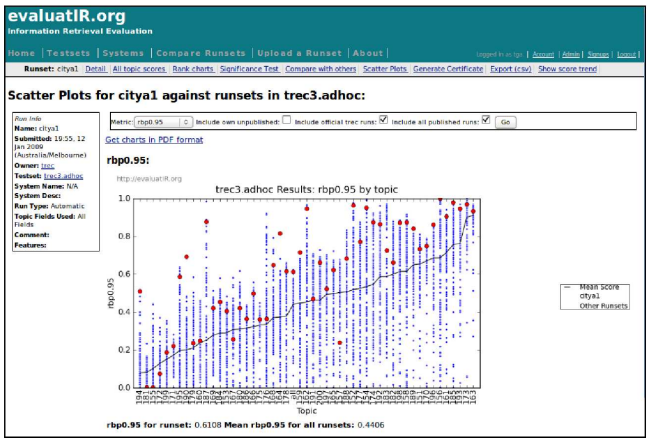


Figure 7: Sample output from the runs database at `www.evaluatIR.org`. This figure is reproduced from a poster paper presentation by Armstrong et al. [2009a].

It is critically important that runs are collected into the repository, not effectiveness scores. Runs are required for the later addition of new effectiveness metrics, and for some statistical tests.

The closure of the TREC Ad-Hoc track was explained by saying "we now have eight years worth of test collections ... sufficient infrastructure exists so that researchers can pursue their investigations independently, and thereby free TREC resources for other tasks" [Voorhees and Harman, 1999]. In part, the public database we propose would extend the life of the valuable TREC resources, by making the reference set of runs dynamic rather than static.

We anticipate further benefits. A public database would encourage more research groups to select or develop baseline research systems that are genuinely of state of the art effectiveness; if such competitive baseline systems were made available, this would remove the burden on each group of needing to implement their own state of the art system from scratch. Readers and reviewers of papers could easily and transparently assess effectiveness claims using more complete and independent information. It would also greatly simplify longitudinal analysis of effectiveness results.

A risk in this proposal is that the research process might be reduced to a simple contest of obtaining the highest numbers. However, we do not believe that the community would be so facile as to regard effectiveness numbers as being the sole determinant of merit: many other considerations come into play.

We have created such a system, available at `www.evaluatIR.org`, and populated it with a range of TREC data. An overview of the system is also available as a poster presentation [Armstrong et al., 2009a]. Figure 7, taken from that poster, shows one form of the output that is available. In this screenshot, a system is compared against the pool of previous published results for that collection and metric, and its relative position in the ranking is highlighted. We invite other researchers – and referees – to explore this website, and to make use of it at every opportunity.

8. CONCLUSION

Our longitudinal analysis of published IR results in SIGIR and CIKM proceedings from 1998-2008 has uncovered the fact that adhoc retrieval is not measurably improving.

There are many possible explanations for this apparent stagnation. What is surprising is that it appears to have gone largely unnoticed within the IR community. An analysis of the papers surveyed provides several reasons for why this may have happened, including selection of weak baselines that can create an illusion of incremental improvement, and insufficient comparison with previous results.

A central repository of effectiveness results presents a solution to this problem: best known results could be quickly found by authors, and readers and reviewers could more effectively assess claims made in papers. We have created such a system, available at `www.evaluatIR.org` [Armstrong et al., 2009a], which we hope will become a valuable resource for the IR community.

Perhaps most urgently of all, though, we should as a community take stock of the situation we find ourselves in. It may be that significant improvements off weak baselines are meaningful. But continuing indefinitely to provide the same quantum of improvement over the same modest baselines inspires neither confidence in our experimental method nor conviction of the contribution of our research. Indeed, as a concrete challenge, perhaps it is time for us to take on what should be an attainable goal – let us build a public system that matches the BM25 run in the 1994 TREC-3 experiment, and then add to it the fruits of the past fifteen years' research, to form a new baseline against which future effectiveness improvements can be properly measured.

Acknowledgments. This work was supported by the Australian Research Council. The inclusion and format of Table 1 was pertinently suggested by an anonymous referee.

References

T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. EvaluatIR: Measurement and certification of IR systems. In *Proc. 32nd Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, page 834, Boston, USA, July 2009a.

T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Has adhoc retrieval improved since 1994? In *Proc. 32nd Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 25–26, Boston, USA, July 2009b.

C. Buckley. The SMART project at TREC. In Voorhees and Harman [2005], chapter 13.

C. L. A. Clarke, F. Scholer, and I. Soboroff. The TREC 2005 terabyte track. In *Proc. TREC-14*, November 2005. NIST Special Publication 500-266.

C. W. Cleverdon. The Cranfield tests on index language devices. *Aslib Proceedings*, 19:173–192, 1967.

C. W. Cleverdon. The significance of the Cranfield tests on index languages. In *Proc. 14th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 3–12, Chicago, Illinois, United States, 1991.

A. P. de Vries and T. Roelleke. Revelance information: a loss of entropy but a gain for IDF? In *Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 282–289, Salvador, Brazil, August 2005.

H. Fang and C. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *Proc. 29th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 115–122, Seattle, USA, August 2006.

D. K. Harman. Overview of the second text REtrieval conf. (TREC-2). In *Proc. TREC-2*, September 1993. NIST Special Publication 500-215.

D. Hiemstra, S. E. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proc. 27th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 178–185, Sheffield, United Kingdom, August 2004.

V. Lavrenko and W. B. Croft. Relevance based language models. In *Proc. 24th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 120–127, New Orleans, LA, USA, September 2001.

S. Liu, C. Yu, and W. Meng. Word sense disambiguation in queries. In *Proc. 14th ACM Int. Conf. on Information and Knowledge Management*, pages 525–532, Bremen, Germany, November 2005.

T. R. Lynam, C. Buckley, C. L. A. Clarke, and G. V. Cormack. A multi-system analysis of document and term selection for blind feedback. In *Proc. 13th ACM Int. Conf. on Information and Knowledge Management*, pages 261–269, Washington, D.C., USA, November 2004.

D. Metzler and W. B. Croft. A Markov random field model for term dependencies. In *Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 472–479, Salvador, Brazil, August 2005.

D. Metzler, T. Strohman, Y. Zhou, and W. B. Croft. Indri at TREC 2005: Terabyte track. In *Proc. TREC-14*, November 2005. NIST Special Publication 500-266.

M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 206–214, Melbourne, Australia, August 1998.

A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Transactions on Information Systems*, 27 (1):1–27, 2008.

M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proc. 28th Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 162–169, Salvador, Brazil, August 2005.

M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. 16th ACM Int. Conf. on Information and Knowledge Management*, pages 623–632, Lisboa, Portugal, November 2007.

E. M. Voorhees and D. K. Harman. Overview of the eighth text retrieval conf. In *Proc. TREC-8*, November 1999. NIST Special Publication 500-246.

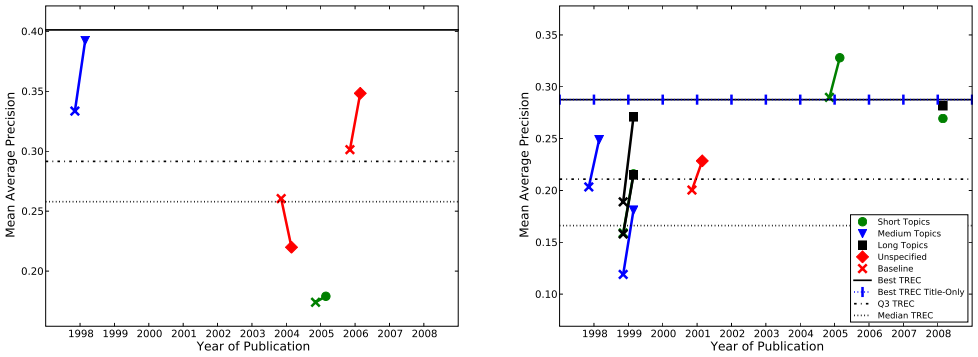
E. M. Voorhees and D. K. Harman, editors. *TREC: Experiment and Evaluation in Information Retrieval*. The MIT Press, 2005.

W. Webber, A. Moffat, and J. Zobel. Score standardization for inter-collection comparison of retrieval systems. In *Proc. 31st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 51–58, Singapore, Singapore, July 2008.

C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2):179–214, 2004.

W. Zhang, S. Liu, C. Yu, C. Sun, F. Liu, and W. Meng. Recognition and classification of noun phrases in queries for effective retrieval. In *Proc. 16th ACM Int. Conf. on Information and Knowledge Management*, pages 711–720, Lisboa, Portugal, November 2007.

J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 307–314, Melbourne, Australia, August 1998.



(a) TREC-3 Ad-Hoc ($n = 4$)

(b) TREC-6 Ad-Hoc ($n = 9$)

(c) TREC-2003 Robust ($n = 6$)

(d) TREC-2004 Robust ($n = 15$)

(e) TREC-2005 Robust ($n = 3$)

(f) TREC-9 Web ($n = 12$)

(g) TREC-2001 Web ($n = 11$)

(h) TREC-9+2001 Web ($n = 11$)

Figure 8: Published MAP scores for different TREC collections used in papers published in SIGIR and CIKM. For the composite TREC-9+2001 Web collection there is no set of original submissions to derive quartiles from.