Examining Additivity and Weak Baselines

Sadegh Kharazmi, RMIT University & NICTA
Falk Scholer, RMIT University
David Vallet, Google
Mark Sanderson, RMIT University

We present a study of which baseline to use when testing a new retrieval technique. In contrast to past work, we show that measuring a statistically significant improvement over a weak baseline is not a good predictor of whether a similar improvement will be measured on a strong baseline. Indeed, sometimes strong baselines are made worse when a new technique is applied. We investigate whether conducting comparisons against a range of weaker baselines can increase confidence that an observed effect will also show improvements on a stronger baseline. Our results indicate that this is not the case – at best, testing against a range of baselines means that an experimenter can be more confident that the new technique is unlikely to significantly harm a strong baseline. Examining recent past work, we present evidence that the IR community continues to test against weak baselines. This is unfortunate, as in the light of our experiments we conclude that the only way to be confident that a new technique is a contribution is to compare it against, nothing less than the state of the art.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Baselines, information retrieval, evaluation

ACM Reference Format:
ACM Trans. Inf. Syst. 0, 0, Article 0 (2016), 18 pages.
DOI: http://dx.doi.org/10.1145/2882782

1. INTRODUCTION

Much of the field of Information Retrieval (IR) is an empirical subject, where the value of novel *techniques* is demonstrated by measuring statistically significant improvements over one or more baselines. This approach was scrutinized by Armstrong et al. [2009], who extensively examined published research showing that in most cases, chosen baselines were substantially less effective (*weaker*) than the state of the art at the time. Worse, the reported improvements from new techniques often resulted in systems that very rarely beat the state of the art.

Armstrong et al. asked if a technique was shown to significantly improve a weak baseline, did that result *predict* that an improvement over a strong baseline would also occur? They conducted an experiment that tested this question measuring what they called a technique's *additivity*. Using the Indri search engine [Strohman et al. 2005], Armstrong et al. tested the additivity of common retrieval techniques (e.g. stop words, stemming, query expansion, query word proximity, phrase matching, and term

Author's addresses: S. Kharazmi, Computer Science, School of Science, RMIT University, Melbourne, Australia; F. Scholer, Computer Science, School of Science, RMIT University, Melbourne, Australia; D. Vallet, Google, Sydney, Australia. M. Sanderson, Computer Science, School of Science, RMIT University, Melbourne, Australia;

DOI: http://dx.doi.org/10.1145/2882782

smoothing). While the improvement gained by each technique was found to be additive on average, the benefit to effectiveness of a single technique was related to which other techniques were enabled at the time. The researchers stated that "*improvements are additive on average, [however] they are not additive always, and additivity needs to be confirmed in individual cases*". The authors suggested testing a new technique "*against a range of configurations*". Because additivity did not always occur, Armstrong et al. [2009] concluded "*we question the value of achieving even a statistically significant result over a weak baseline*".

Armstrong et al.'s additivity experiments tested classic retrieval techniques applied to a single retrieval system. To the best of our knowledge, there are very few other studies testing additivity. Therefore we conduct additivity experiments on newer retrieval techniques against a range of baseline systems. We asked the following research questions.

(1) Do more recently developed retrieval techniques show similar additivity effects to those found by Armstrong et al.?
(2) If improvements of a technique over a single weak baseline do not accurately predict what will occur when the technique is applied to a strong baseline, can the use of multiple weak baselines improve the accuracy of such predictions?
(3) In the light of Armstrong et al.'s survey of past papers, are stronger baselines now being used in the research community?

We carry out two studies of a more recent retrieval techniques: search result diversification; and ranking of microblog content. The codes and run data for our experiments is available.[1] In the paper, we examine past work; describe the setup, results, and analysis of the two studies; then we draw conclusions and describe possible further work.

2. PAST WORK

The work of Armstrong et al. has had a substantial and on-going impact in the field of IR. Beyond its many citations, the paper acted as a catalyst to encourage more rigorous evaluation. Subsequently, workshops (e.g. RIGOR at SIGIR 2015), and even whole conference sessions on reproducible IR have been held [Ferro and Silvello 2015; Hagen et al. 2015; Rao et al. 2015].

The paper contains multiple novel experimental results, leading to a number of key conclusions around which we structure the survey of past work. The main conclusions were ad hoc retrieval effectiveness has plateaued; weak baselines are commonly used by researchers; and improvements are sometimes, but not always, additive.

2.1. Ad hoc retrieval effectiveness has plateaued

Probably, the best known aspect of Armstrong et al. study comes from the authors' analysis of 106 past papers published in key IR conferences. The main results of this analysis was reported in a series of scatter plots in the fourth and eighth figures of their paper, the first of which is reproduced in Figure 1.

Results from the past papers are graphed as a cross and point connected by a line. The cross indicates the effectiveness of a baseline system, the point indicates the change in system effectiveness resulting from the application of a retrieval technique. Horizontal lines indicate the effectiveness of state of the art systems. The common focus of the graphs are the upper black lines, which represent the highest effectiveness score recorded by TREC organizers in the year that the collection under study was released. The line is almost never crossed suggesting little or no improvement.

[1] http://www.rmit-ir.org/index.php/resources/tois2016

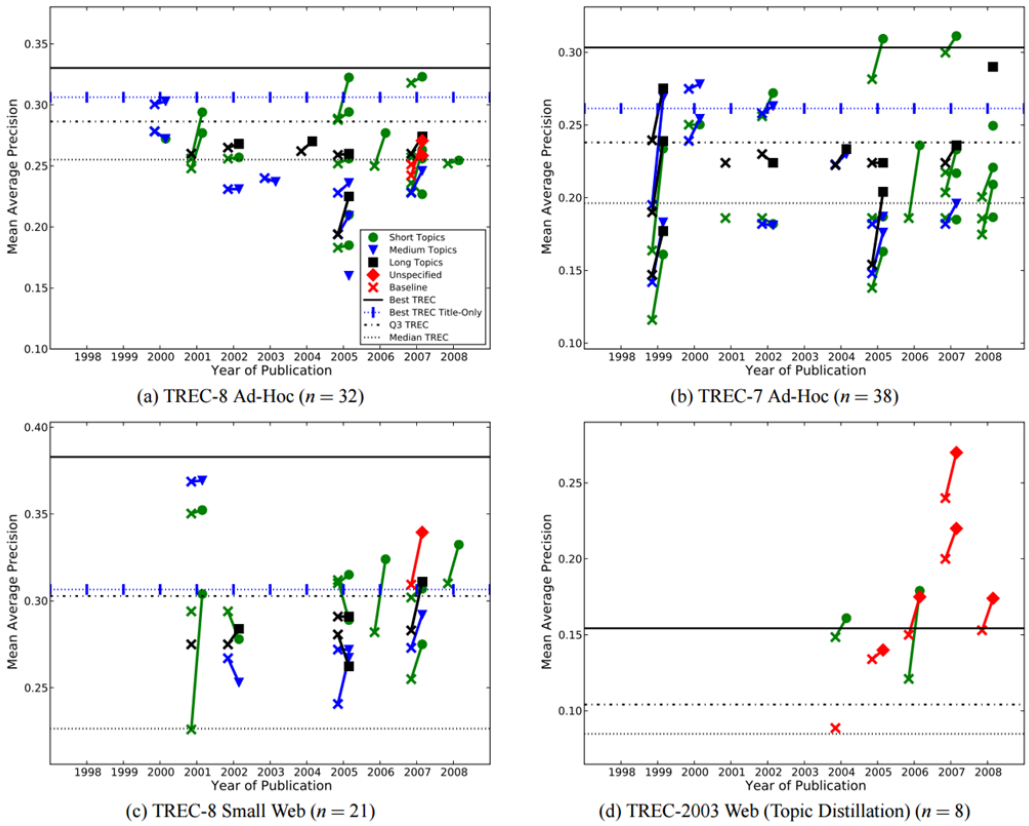


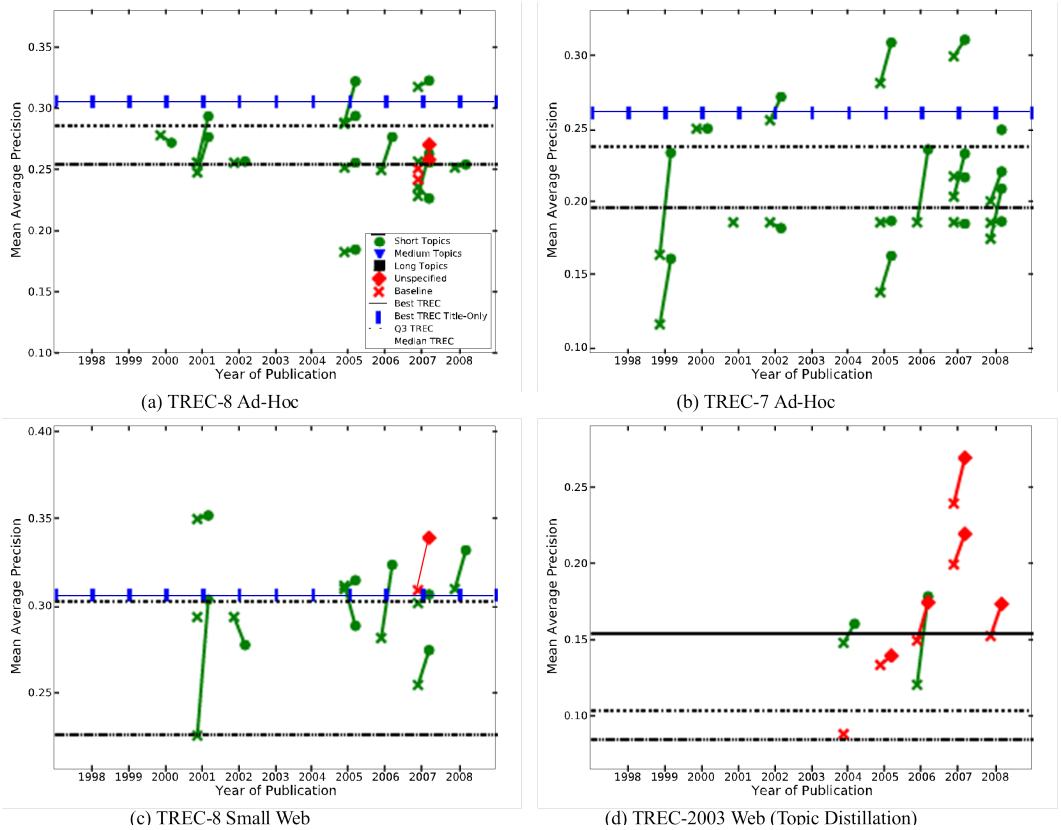
Fig. 1. Reproduction of Figure 4, taken from Armstrong et al. [2009].

The result is regularly cited – [Bodoff 2013; Cummins et al. 2011; Fang et al. 2014; Fuhr 2012; Leveling et al. 2012; Mayr et al. 2014; Newman et al. 2010; Puurula 2013; Said and Bellogín 2014; Sakai and Lin 2010; Stegmaier et al. 2014; Trotman et al. 2014; Ye et al. 2012] – with authors sometimes augmenting the result to claim that Armstrong et al. showed *no* improvements had occurred in ad hoc IR.

However, there is a danger that the graphs can be mis-interpreted as they plot multiple experimental conditions. Some comparisons are less meaningful than others. Armstrong et al. grouped their results by the length of topic (e.g. short, medium, and long) used in the published experiments. Although it is not explicitly stated in the paper, we believe that the three groups correspond to combinations of TREC topic sections: title, title+description, and title+description+narrative. Note, in some published papers, the type of topic length was not stated, we assume such topics are title only.

Comparing an experimental result using title only topics with a result using title+description+narrative (commonly the topic used to achieve the score indicated by the upper black line) is perhaps not the most valuable comparison to make. Potentially more valuable is a comparison with the highest recorded score for title (short) topics: the dashed line with tick marks.

When only title topics are considered (see Figure 2), we see that while it remains true that many retrieval experiments published in top IR conferences used weak baselines, not all experiments used such baselines. Stronger baselines were sometimes used and across all collections, the state of the art score recorded by TREC was beaten more than



(a) TREC-8 Ad-Hoc (b) TREC-7 Ad-Hoc

(c) TREC-8 Small Web (d) TREC-2003 Web (Topic Distillation)

Fig. 2. Revised reproduction of Figure 4 taken from Armstrong et al. [2009].

once. For results using the collections shown in the bottom two graphs of Figure 2, the majority of published results beat the state of the art at the time. Note, further results covering additional test collections were shown in the eighth figure of Armstrong et al. [2009] where a similar misinterpretation could be made.

The work of Armstrong et al. tackled an important and overlooked topic. The novel analysis of past work showed that the IR community was willing to publish results that tested techniques, against sometimes very weak baselines. However, over the decade of research examined by the authors, improvements over strong baselines using title only topics *were* shown, and on some collections were shown regularly.

Subsequent work. A series of papers were published to examine the plateauing of effectiveness apparently shown by Armstrong et al..

Trotman and Keeler [2011] examined if inconsistency in the way that test collection assessors marked documents as relevant caused there to be an upper bound on measurement accuracy. Testing on TREC-4 and TREC-6 collections as well as a small selection of INEX collections, the authors found some support for their hypothesis. However the experiments in this short paper were not extensive.

Taking a different approach Cummins et al. [2011] claimed, through oracle experiments, that there was still much improvement possible in ad hoc retrieval. The authors searched for patterns of terms that formed ideal ad hoc topics. They found that queries could be located that nearly doubled MAP scores when compared to a baseline system.

More recently Trotman et al. [2014] examined a number of recently published enhancements to the BM25 ranking function [Robertson et al. 1995]. Trotman et al. showed that the way the parameters of ranking functions were tuned had an impact on retrieval effectiveness. When each variant of BM25 was tuned optimally for a particular collection, comparisons between the different functions were more complex and conclusions were harder to draw from results.

Demonstrating that it is possible to beat state of the art ad hoc systems, Huston and Croft [2014] examined improvements through exploitation of topic-term dependencies. The detailed review of dependency models considered optimal parameter tuning and testing on multiple collections (Robust, Gov2, Clueweb-09). Huston and Croft showed that one could improve on state of the art rankers. The work also determined which dependency model was generally more effective than the others.

Although it is hard to beat state of the art ad hoc rankers, it would appear that the assertion that ad hoc IR has not improved in the last several years is not as well supported as perhaps was thought.

2.2. Weak baselines are commonly used

Although evaluation of IR systems has been a topic of interest for decades and test collections proliferated, the practice of recording the highest effectiveness score on them did not. For example, Spärck Jones and van Rijsbergen [1976] surveyed and tabulated around thirty contemporary test collections, without mentioning the best retrieval techniques or highest recorded score. Sanderson [2010] similarly tabulated collections, but without details of scores. In contrast, Downie et al. [2010] surveying test collections from the music IR evaluation campaign MIREX, listing the best score for each effectiveness measure on each collection (see table 3 in that paper).

The organizers of TREC annually record the score of the most effective technique on a particular collection Voorhees and Harman [2005], but Armstrong et al. showed that few researchers or reviewers pay sufficient attention to such scores. Each year, new collections and topics are created by TREC track organizers so as to present a new research challenge and to cumulatively create a large test data set. Because the variation in effectiveness across topics is high [Bodoff and Li 2007] – as are the interaction effects between systems, topics, and documents – it is, in effect, invalid to compare techniques tested on different years of a track.[2]

Other fields with empirical evaluation methods charted the progress not just of individual systems, but of their wider research community. For example in the speech recognition community [Huang et al. 2014], steady improvements in accuracy were charted over many years (see Figure 1 in that paper). In the Music IR community, using data from the MIREX evaluation, similar graphs were produced [Schedl et al. 2014] (see Figure 5.1 in that paper). Of note in this later analysis, it was found that researchers were failing to improve on certain MIREX tasks. This plateauing was referred to as a glass ceiling by Downie [2008], who contended "*that this evidence has been one of the most important contributions of MIREX to MIR research*". The impact of identifying the ceiling was that researchers in the field chose to move onto new research challenges.

We can only speculate if "weak baseline papers" such as those identified by Armstrong et al. would have been accepted had the state of the art been tracked more diligently by IR researchers.

[2]See Webber et al. [2008], describing an attempt to estimate such comparisons across topic sets.

To the best of our knowledge Armstrong et al. is the only study of the strength of baselines in use by researchers. We searched extensively – including examining the 76 papers that cite Armstrong et al.[3] – and found no similar study in the last six years.

2.3. Improvements are sometimes additive

As with the work on strength of baselines, beyond Armstrong et al.'s experiment on additivity, there is little work published in the last six years which examines the additivity of methods across different baselines.

Trotman et al. [2014] in their study comparing variants of BM25, also explored the value of applying different retrieval techniques such as stemming, stop word removal, and pseudo-relevance feedback across a number of rankers and document collections. As with Armstrong et al., Trotman et al. found that the impact on effectiveness from the inclusion of the techniques was generally additive; though unlike Armstrong et al. the use of stop words was not found to help.

One might ask why is additivity important? A new technique should be tested against the state of the art. However, there has been a growing concern in the research community that state of the art methods are not always accessible due to commercial interests or a lack of access to proprietary data (see [Callan and Moffat 2012]). Consequently, it may be necessary to test techniques on sub-optimal baselines. Determining if one can reliably conduct such testing is a question that we study here.

3. DIVERSIFICATION

The core of our work was to examine additivity in the context of novel retrieval techniques applied to ad hoc search. We examined two such techniques: search result diversification (this section) and syntactic representations of microblog content (section 4).

Search result diversification attempts to retrieve a set of documents that reflect the sub-topics of a query. The study of diversification was not extensive until after Armstrong et al. published their work. Consequently, this is a topic of interest for our work. We first briefly describe diversification approaches, methods and data, and then detail the results of our experiments.

3.1. Diversification Approaches

Santos et al. [2012] categorized diversification techniques as *implicit* or *explicit*: implicit techniques model diversity just from the content of retrieved documents [Carbonell and Goldstein 1998; Sanner et al. 2011; Wang and Zhu 2009; Zhai and Lafferty 2006], while explicit approaches model aspects of a query using external evidence such as query logs, thesauri, or Wikipedia [Agrawal et al. 2009; Santos et al. 2012].

Most implicit approaches are based on a greedy approximation, which aims to reduce redundancy in retrieved documents with respect to their content. The focus of these approaches is on promoting novelty in the ranking. We used three implicit diversification approaches in our experiments.

— Maximal Marginal Relevance (MMR) [Carbonell and Goldstein 1998], builds a diverse set of results (S) incrementally, from an ad hoc run (R) using a greedy approach in which, in each iteration, the most novel and relevant document is selected. Novelty is defined as the mean content-based dissimilarity between the candidate document and the already selected documents in S.
— Modern Portfolio Theory (MPT) [Wang and Zhu 2009], considers the uncertainty associated with the relevance of documents to a query. The theory could be observed as

[3]Number taken from Google Scholar, accessed July 2015: https://scholar.google.com.au/scholar?cites=15655226943540577380

a process of selecting a set of documents which reduce the uncertainty about document relevance. The relevance of a ranked list of documents should be maximized, and the variance minimized.
— Facility Location Analysis (FLA) [Zuccon et al. 2012], here, an approach from the field of Operations Research is applied. Documents are viewed as facilities, rank position of those documents are locations. In FLA desirable facilities should be located close by and, so-called, obnoxious facilities (i.e. documents covering the same subtopic) should be dispersed as far as possible from each other.

Explicit diversification approaches directly model the subtopics of a query [Santos et al. 2010]. We consider three approaches.

— Explicit Query Aspect Diversification ($xQuAD$) [Santos et al. 2010], ranks documents based on the number of query subtopics that occur in a document, the overall relevance, and a document's novelty relative to the documents ranked higher.
— Intent-Aware Selection ($IASelect$) [Agrawal et al. 2009] diversifies results based on a taxonomy, such as the Open Directory Project (ODP), to indicate which of the possible intents (subtopics) a query could cover.
— Relevance Based xQuAD ($xQuADRel$) [Vargas et al. 2012], here a formal relevance model is defined, which is applied to $xQuAD$.

The six diversification techniques were implemented. The code for $xQuADRel$ was provided by Vargas et al.; the other approaches were implemented as described in their respective papers. Diversification was applied to the top 100 documents of each ad hoc run. For validation, we compared the effectiveness of our implementations with published results. While identical effectiveness is typically not possible when source code is not available, the general effectiveness scores and trends were consistent with those reported.

Subtopics. The diversity approaches $xQuAD$, $xQuADRel$, and $IASelect$ need to be parameterized with query subtopics. Two subtopic definitions were used: the TREC Web Track subtopics; and those derived from the ODP, using TextWise[4] services. The TREC subtopics were based on information taken from relevance judgments, they represented an upper bound on the effectiveness of these approaches due to subtopic coverage. The ODP subtopics represented a reasonable but imperfect set of subtopics. The two sources are indicated using the subscript labels TREC and ODP in the results.

Training and Tuning. All parameters in the diversification approaches were optimized using 5-fold cross-validation. Parameter values were considered in the range 0.0–1.0, in increments of 0.1. For $xQuAD_{ODP}$, the best λ value was 0.8, while for $xQuAD_{TREC}$ it was 0.9. For MMR the optimal λ was 0.7. The $xQuADRel$ technique has an additional parameter, $P(stop|r)$, which we set to the suggested default of 1. The other approaches required no training.

3.2. Test Collection and Runs

The six techniques re-rank a set of retrieved documents independently of the initial ranking algorithm. Therefore, we decided to apply the techniques to the runs submitted to a TREC ad hoc retrieval track, which would simulate applying the diversity techniques to a wide range of baselines.

We chose the TREC Web Track, 2009–2011, which had separate ad hoc and diversification submitted for the same topics. Therefore we had access to ad hoc runs for which

[4]http://www.textwise.com/

Table I. IR technique count used across ad hoc runs

Technique	Count
Well known ranking methods with parameter variation	52
Document/Collection feature (e.g. fields, anchor text)	19
External resources (e.g. Wikipedia)	18
Relevance feedback	10
Query expansion	12
Learning to rank methods	20

Note, categories were not mutually exclusive.

there were also diverse relevance judgments. We used the automatic ad hoc runs over the ClueWeb B collection.

In line with the methodology of previous research analyzing ad hoc run data [Sanderson et al. 2012; Voorhees and Buckley 2002], the bottom 25% of submitted runs were removed in an attempt to filter out problematic submissions that could produce invalid comparisons. Across the three years of the track, this left 71 runs. Each run recorded documents retrieved for fifty topics, corresponding to one of the years of the track. Because our focus is on the way that the effectiveness of a run is improved by diversification and not on variations across the years of the collection, we present results aggregated across all three years.

The systems that generated the ad hoc runs were checked to ensure that they did not employ diversification techniques. This was verified by manually inspecting 52 track reports[5] submitted by TREC participants to describe the techniques employed by the systems that generated the 71 runs. None of the reports described use of diversification techniques. The 2009–2011 ad hoc runs were therefore appropriate baselines for our study.

We also examined the reports to understand the techniques that were used, shown in Table I. In addition to using well known retrieval techniques, many participants exploited query expansion, or made use of additional information such as anchor text or other external resources.

Evaluation Metrics. The primary effectiveness measures for the evaluation of diversified runs in the TREC Web track were $\alpha-$nDCG [Clarke et al. 2008] and IA-ERR [Chapelle et al. 2009] at cut-off 20. For $\alpha-$nDCG, the parameter α, was set to the default 0.5. We report the same measures in our experiments.

3.3. Improvements Over Weak and Strong Baselines

We compared the effectiveness of the 71 ad hoc runs before and after a chosen diversity technique was applied. The results are shown in Figure 3. The top two rows of graphs show techniques that use subtopics (*xQuAD*, *xQuADRel*, *IASelect*); in the top row, the official TREC subtopics are used, while the middle row shows the same approaches using the ODP subtopics. The bottom row shows techniques that do not use subtopics (*MMR*, *MPT*, *FLA+MPT*). The following trends can be observed.

(1) In the top row (TREC subtopics), the effectiveness of weaker baselines was almost always improved by diversity. For stronger baselines, there was almost never a significant improvement, but no significant degradation. For $IASelect_{TREC}$ a single strongest baseline was significantly improved beyond the highest score of any baseline.
(2) For the middle row (ODP subtopics), effectiveness improvements were rarely significant, and for two approaches (*xQuARel* and *IASelect*) significant degradation was

[5]Not every report was available in the online TREC proceedings

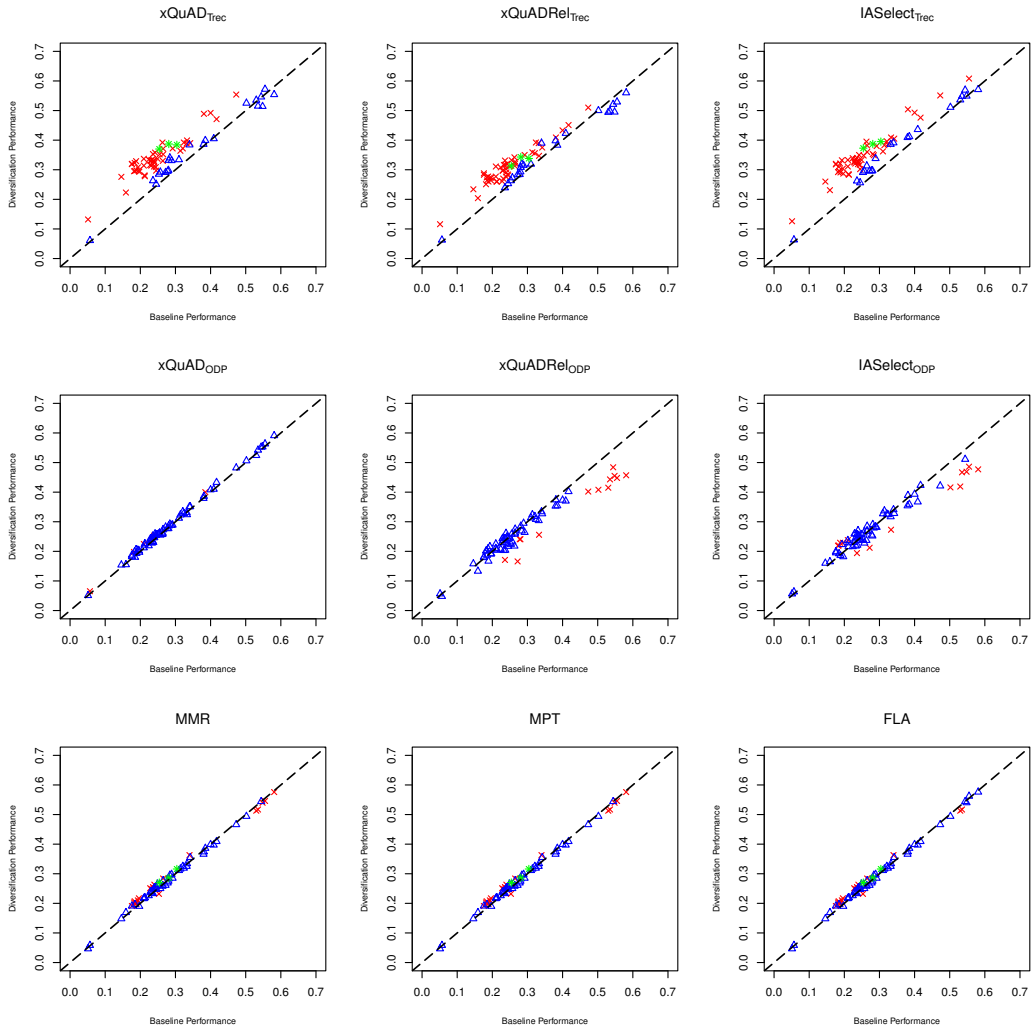


Fig. 3. Scatter plots showing the impact of nine diversity techniques. Each point shows the effectiveness of a technique applied to an ad hoc baseline run. The weaker baselines were plotted to the left of each graph, the stronger baselines to the right. The scale of the x- and y-axis are both measured using $\alpha-$nDCG@20. Points plotted above the 45-degree line represent the technique improving a baseline, plots below show degradation. A point plotted as a red cross indicates a significant difference (positive or negative) relative to the baseline. A blue triangle indicates no significant difference. A paired two-tailed t-test was used. Green stars show effectiveness of the three additional baseline runs using the Indri retrieval system configured to use different rankers (note, this was only conducted on the top and bottom sets of techniques).

commonly measured on the stronger baselines. For $xQuAD_{ODP}$ and $IASelect_{ODP}$, a few baselines were significantly improved, but the majority of changes were not significant or significantly worse.

(3) For the techniques that do not use subtopics (the bottom row), non-significant changes were generally found, with some small significant improvements over weaker baselines, and small significant degradations for stronger baselines.

Table II. Range of scores used for levels

Level	$\alpha-$nDCG@20	IA-ERR@20
Weak	≤ 0.23	≤ 0.18
Medium	> 0.23 & ≤ 0.41	> 0.18 & ≤ 0.33
Strong	> 0.41	> 0.33

Note, as a sanity check of our experiments, which augment existing ad hoc runs, we tested applying diversity to an actual IR system: INDRI. It was set up in three configurations: Okapi BM25 weighting, language modeling, and sequential dependency modeling. These experiments were conducted on all 148 topics of the combined web collections and were tested on the top and bottom rows of techniques. The results were plotted as green stars in Figure 3. As can be seen, the $\alpha-$nDCG@20 scores from INDRI and the improvements derived from application of diversity were in line with what we found with the ad hoc run experiments.

Oracle Runs. A possible explanation for the failure to improve over strong ad hoc runs is a saturation effect where the ad hoc run is already diverse, and so little or no further improvement is possible by applying a subsequent diversification technique. Such an effect was reported by Santos et al. [2012]. We therefore generated a set of oracle runs to identify the upper bound of effectiveness in the TREC ad hoc runs.

The best possible diversified rankings were created using the diversity relevance judgments provided by TREC to re-order the top 100 retrieved documents of each ad hoc run. Using the subtopic information in the TREC relevant judgments, documents were re-ordered so that $\alpha-$nDCG@20 was maximized. In effect, the relevant documents in the top 100 were placed at the top of the rankings with subtopics distributed evenly. It was found that the oracle runs were on average 24% better than the diversified runs plotted in Figure 3. The improvement was found for both diversified weak and strong ad hoc runs.

It would appear that saturation is not the reason for lack of improvement by diversification techniques, as there is still room for improvement even for strong baselines.

3.4. Defining and Examining Strong and Weak Baselines

We grouped ad hoc runs into one of three levels: weak, medium, and strong. To do the grouping, for each metric, the minimum and maximum run scores were noted and three equally sized levels were formed within the range. The effectiveness score boundaries are defined in Table II. Since the ranges were partitioned by score, different numbers of runs fell into each level; this is shown in Table III.

The retrieval techniques from the 52 TREC reports (tabulated in Table I) are split into three analysis levels, as shown in Table IV. As can be seen, learning to rank methods, along with a range of web specific text features characterized the strong baselines. By and large, using classic "vanilla" IR techniques without enhancements characterize the weak baselines.

Echoing Armstrong et al.'s review of papers, we manually searched for diversification related publications appearing in SIGIR in 2010–2012. From 51 full length papers, short papers, posters, and workshop reports collected, we identified seven papers [Dang and Croft 2012; He et al. 2012; Santos et al. 2011a,b,c; Vallet and Castells 2012; Vargas et al. 2012] that introduced a new diversification technique or a way of improving the effectiveness of existing approaches, and evaluated these using the TREC Web Track test collection.

Of the baselines used in these publications, 44.4% occurred in the weak level of our earlier defined ranges, while 55.6% occurred in the medium levels. No reported baselines were in the strong level.

Table III. Number of baselines in each level

Level	$\alpha-$nDCG@20	IA-ERR@20
Weak	19	42
Medium	42	21
Strong	10	8

Table IV. Techniques used in the ad hoc runs broken into three effectiveness levels

Level	Year	Description
Weak	2009,2010	— Default lemur ranking (language modeling) — Different relevance models and merging the results from them — Combination of query likelihood, dependence model, relevance model, and domain trust prior — Different variations of IDF — Including semantic term matching derived from collections — Default OKAPI BM25 ranking — Mixture of document scores, anchor-text, and spam score
Medium	2009, 2010	— Different retrieval functions (LM, OKAPIi, DFR)+spam filtering with different thresholds — Query expansion from different sources (collection, Wikipedia) — Extending BM25 (taking into account different features like query term proximity) — Probabilistic Data Fusion — Semantic term matching derived from web search engines — Enhancement of Lucene ranking using different indexes and domain duplication removal — Machine learned ranking using a wide range of features — Weighted sequential dependence model — Optimized retrieval to maximized precision — Combination of anchor text, document score, and spam filtering — Phrase search based on n-grams
Strong	2010, 2011	— Machine learned ranking using a pool of features — Combination of anchor text, full text, and title combined using linear smoothing with a spam prior — Anchor text + merging results from different sources

3.5. Do Improvements on Weak Baselines Predict Improvements on Strong Baselines?

The percentage of runs that were significantly improved or degraded by the application of a diversity technique are shown in Table V. While the trends for both metrics were similar, for IA-ERR@20, the mean percentage of significant improvements for all three levels was lower than for $\alpha-$nDCG@20. The opposite held for significant degradation.

Overall, techniques not using subtopics consistently made strong baselines significantly worse. The subtopic-based techniques were more successful, though only those using the precise TREC sub-topic definitions significantly improved the strongest baselines, while making none of the other baselines significantly worse.

We considered the relationship between the percentage of weak or medium baselines that were significantly improved, and the percentage of strong baselines significantly degraded. Figure 4 plots this relationship for the nine diversification techniques, when using weak baselines. Correlations between 'strong baselines significantly degraded' and 'weak baselines significantly improved' for $\alpha-$nDCG@20 and IA-ERR@20 were -0.72 and -0.75 respectively. Although the number of data points was small, both correlations were significant ($p < 0.05$). However, further testing generating more data points is necessary for us to be more confident of this result.

3.6. Diversification Methods as a Baseline

The experiments so far simulated a researcher trying a new diversification technique for the first time, and comparing effectiveness to a baseline ad hoc system. More commonly, researchers report enhancements to existing techniques and compare improvements over previous versions. We compared the effectiveness of one diversification

Table V. Percentage of runs that were statistically significantly improved or degraded for each diversification technique, grouped by level.

Level	Diversity technique	Improved		Degraded	
		$\alpha-$nDCG@20	IA-ERR@20	$\alpha-$nDCG@20	IA-ERR@20
Weak	$xQuAD_{TREC}$	94.7	80.9	0	0
	$xQuADRel_{TREC}$	89.4	73.8	0	0
	$IASelect_{TREC}$	94.7	83.3	0	0
	$xQuAD_{ODP}$	26.3	11.9	0	0
	$xQuADRel_{ODP}$	21.0	7.9	0	4.7
	$IASelect_{ODP}$	26.3	21.4	0	0
	MMR	0	0	5.2	4.7
	MPT	31.5	21.4	0	2.3
	FLA	42.1	26.1	0	0
Medium	$xQuAD_{TREC}$	59.5	14.2	0	0
	$xQuADRel_{TREC}$	45.2	9.5	0	0
	$IASelect_{TREC}$	69.0	42.8	0	0
	$xQuAD_{ODP}$	4.7	0	0	0
	$xQuADRel_{ODP}$	2.3	0	3.5	14.2
	$IASelect_{ODP}$	0	0	7.1	4.7
	MMR	2.3	0	2.3	9.5
	MPT	11.9	0	2.3	9.5
	FLA	16.6	9.5	2.3	4.7
Strong	$xQuAD_{TREC}$	10	0	0	0
	$xQuADRel_{TREC}$	0	0	0	0
	$IASelect_{TREC}$	30	12.5	0	0
	$xQuAD_{ODP}$	0	0	0	0
	$xQuADRel_{ODP}$	0	0	80	87.5
	$IASelect_{ODP}$	0	0	60	50
	MMR	0	0	40	62.5
	MPT	0	0	40	51
	FLA	0	0	20	25

Table VI. Percentage of baseline runs (column) significantly improved ($\alpha-$nDCG@20) by a new technique (row)

New Technique \ Baseline	$xQuAD_T$	$xQuADRel_T$	$IASelect_T$	$xQuAD_O$	$xQuADRel_O$	$IASelect_O$	MMR	MPT	FLA
$xQuAD_{TREC}$	-	5.6	5.6	97.1	95.7	97.1	98.5	98.5	84.5
$xQuADRel_{TREC}$	8.45	-	8.4	94.3	98.5	95.7	98.5	95.7	83
$IASelect_{TREC}$	5.6	7	-	98.5	97.1	98.5	100	97.1	87.3
$xQuAD_{ODP}$	0	0	0	-	2.8	14	56.3	16.9	5.6
$xQuADRel_{ODP}$	0	0	0	14	-	0	54.9	15.4	7.4
$IASelect_{ODP}$	0	0	0	2.87	2.8	-	57.7	18.3	4.2
MMR	0	0	0	0	0	0	-	1.4	0
MPT	0	0	0	4.2	1.4	4.2	46.4	-	5.6
FLA	0	0	0	7.4	9.8	9.8	59.7	35.2	-

technique against another across all 71 runs. Table VI shows the percentage of significant improvements (paired two-tailed t-test) that were observed when comparing one of the nine diversification approaches (row) as a novel approach to another technique (column).

For example, for the row $xQuADRel_{TREC}$ and column MMR, MMR was the baseline technique and $xQuADRel$ with TREC subtopics was the new technique. The later technique achieved significantly better results on 98.5% of the 71 base runs. As can be seen, the strongest baselines (the left-most columns) were again the hardest to beat, in-keeping with the observations of Section 3.3.

4. MICROBLOGS

The next study examined the results of a recent experiment on microblog retrieval, where the authors tabulated extensive result and significance data [Severyn et al. 2014] (see Table 2 of the paper). The experiment tested three variations of a retrieval technique across thirty baselines.

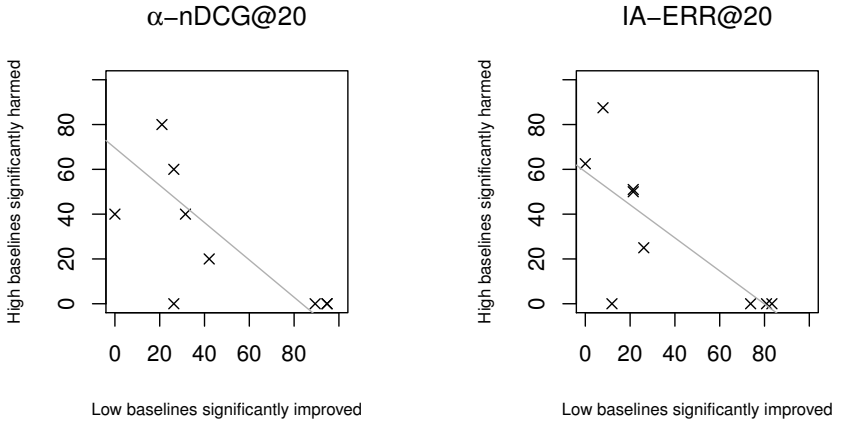


Fig. 4. The percentage of strong baselines that are significantly degraded, versus weak baselines that are significantly improved, using the initial retrieval runs as baselines. A line of best fit is shown. The correlation for $\alpha-$nDCG@20 is -0.72 & IA-ERR@20 is -0.75; both correlations are significant ($p < 0.05$).

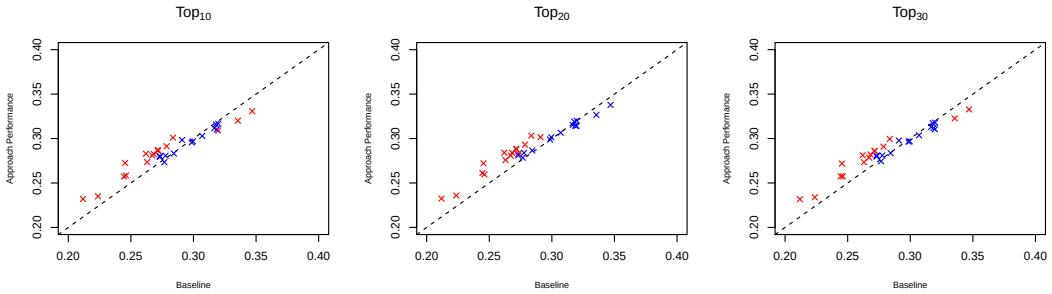


Fig. 5. Three scatter plots showing the effectiveness of three different approaches applied on baseline runs submitted to the Microblog TREC task. Each point shows the effectiveness of an approach applied to a baseline. The weaker baselines are plotted to the left of each graph, the stronger baselines to the right. The scale of the x- and y-axis are both measured using MAP. The 45-degree line indicates the effectiveness of the baseline. Points plotted above the line represent improvement over baseline, plots below show degradation. A red cross indicates a significant difference relative to the baseline. A blue triangle an insignificant change. Significance was measured using a paired two-tailed t-test.

Severyn et al. proposed a new structural representation of queries and the short texts of microblog documents to improve retrieval effectiveness. The authors used a syntactic analysis to transform text into a shallow tree kernel representation. The authors stated that a two level hierarchy was created using lemmas and part-of-speech tags based on a trainable tagger from CMU. The hierarchy was divided using an OpenNLP chunker. Next, a point-wise learning to rank approach was used to match between the queries and documents in their new representational form. The ranker was trained on topics and tweets from the TREC microblog 2011 track. The results, measured on the TREC 2012 collection, showed that using relational syntactic structures improved effectiveness by 5% on average. We use Severyn et al.'s data to study effectiveness and baselines.

Severyn et al. used their technique to re-rank the top-k documents, setting k to 10, 20, and 30. The retrieved documents from thirty runs submitted to the TREC 2012

microblog track were re-ranked and the results were tabulated. This paper is one of the few we found that evaluates against an extensive range of baselines.

We examined all available TREC reports of the thirty runs and found that none of the runs used any form of syntactic analysis when retrieving microblog content. Consequently, we employed the results to study our first research question.

Because there were only three variations of the same technique tested in this work, measuring if beating a weak baseline predicted beating a strong baseline (as examined in Section 3.5) was not possible. However, we plotted the effectiveness of the thirty ad hoc runs before and after each of the three techniques was applied, see Figure 5. Similar to Figure 3 one can observe that the effectiveness of weaker baselines was almost always improved by the chosen technique. However, when the stronger baselines were employed, no significant improvement was observed and for two of the approaches very strong baselines were significantly degraded. These results are directly in line with those of Section 3.3 for diversification approaches.

5. CONCLUSIONS AND FUTURE WORK

This paper examined the following research questions:

(1) Do more recently developed retrieval techniques show similar additivity effects to that found by Armstrong et al.?
(2) If improvements of a technique over a single weak baseline do not accurately predict what will occur when the technique is applied to a strong baseline, can the use of multiple weak baselines improve the accuracy of such predictions?
(3) In the light of Armstrong et al.'s survey of past papers, are stronger baselines now being used in the research community?

Past work suggested that there was a level of additivity in retrieval experiments. If a new technique was created and successfully applied to a weak baseline, there was evidence that the technique was likely to improve a strong baseline. In this paper, a syntax-based text representation technique and nine diversity techniques were tested against, in total, 101 baseline systems. The scatter plots of Figures 3 and 5 showed that while the techniques were often found to improve weak baselines significantly, strong baselines were almost never improved, commonly, they were made significantly worse. In contrast to the past work, additivity almost never occurred.

Both of our analyses showed that a statistically significant improvement over a weak baseline did not predict with any confidence what occurred when that technique was applied to a strong baseline.

Testing on the nine diversity techniques, we next examined if prediction was improved by testing for significance on multiple weak baselines. The data available for this test was not extensive, but a trend was found that observing significant improvement over multiple weak baselines improved prediction of what would occur on a strong baseline. On the question of how many weak baselines needed to be improved, the results shown in Figure 5 suggest that only when $75\% - 100\%$ of the (tens of) weak baselines were significantly improved could one have some confidence about prediction. However, the best prediction available was that a strong baseline would not be made worse. Even with the multiple baselines, additivity over these nine techniques was not found.

We showed that some researchers were tempted to conclude that ad hoc search effectiveness plateaued several years ago, and by implication the plateauing explains the inability to beat strong baselines. However, by re-examining the results of past work, we showed that there was a danger that the results of Armstrong et al. were misinterpreted and that more recent papers showed that ad hoc search can still be made more effective.

One aspect of IR experimental practice that was highlighted in Armstrong et al.'s work that does not appear to have substantially changed in the intervening years is the choice of baseline. An examination of recent past work related to the diversification study indicated that use of strong baselines in published research was still not a common practice. The additivity results of this paper show that testing on strong baselines is critical to demonstrate an advance. However, the use of such strong baselines does not appear to be as common as one might hope given the impact of Armstrong et al.'s paper.

It would appear that additivity is less present than it was in the time of Armstrong et al.'s experiments. As ad hoc systems improve, the quality of testing may also need to improve. Early techniques for improving ad hoc IR may well have had such a substantial positive impact on effectiveness that testing the techniques on any form of baseline was sufficiently safe. As techniques become more complex and improvements are generally smaller, the rigor of testing most likely needs to improve also.

While this analysis chiefly serves to illustrate a problem, rather than presenting a ready solution, we hope that it will lead to further awareness in the IR community regarding the importance of strong baselines.

5.1. Future work

What this study has not examined is what sort of strong baseline should be selected when conducting an experiment. It has been recently argued from a theoretical standpoint that it isn't necessarily the best idea to compare a new technique against the strongest IR system [Bodoff 2013]. In their tutorial on evaluation Metzler and Kurland [2012] described a series of heuristics for choosing a baseline (see slide 43[6]). They suggested selecting a baseline

— that uses the same "*principles*" as the technique being tested;
— that sheds light on the hypothesis represented by the technique;
— that "*stays within the same framework*" as the IR system on which the technique is implemented on;
— that is strong.

There is, to the best of our knowledge, little or no research that tests these principles except the last. Examining baseline selection in more detail is a topic that will be explored in future work.

We also suggest that it might be time to repeat Armstrong et al.'s survey of baseline choices to better understand how much the IR community changed in the light of that paper.

ACKNOWLEDGMENTS

This work was supported by the Australian Research Council's Discovery Projects scheme (DP130104007) and by a PhD scholarship supported by NICTA. We thank the reviewers' invaluable comments on this paper as well as a past version of this paper. We also thank Jamie Callan and Alistair Moffat for early input and discussion.

REFERENCES

R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. 2009. Diversifying search results. In *Proc. WSDM*. ACM, 5–14.

[6]http://iew3.technion.ac.il/~kurland/sigir12-tutorial.pdf

Timothy G Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In *Proc. CIKM*. ACM, 601–610.

David Bodoff. 2013. Fuhr's challenge: conceptual research, or bust. In *ACM SIGIR Forum*, Vol. 47. ACM, 3–16.

D. Bodoff and P. Li. 2007. *Test theory for assessing IR test collections*. ACM New York, NY, USA, 367374.

Jamie Callan and Alistair Moffat. 2012. Panel on use of proprietary data. In *ACM SIGIR Forum*, Vol. 46. ACM, 10–18.

J. Carbonell and J. Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proc. SIGIR*. ACM, 335–336.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proc. CIKM*. ACM, 621–630.

C.L.A. Clarke, M. Kolla, G.V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proc. SIGIR*. ACM, 659–666.

Ronan Cummins, Mounia Lalmas, and Colm O'Riordan. 2011. The limits of retrieval effectiveness. In *Advances in Information Retrieval*. Springer, 277–282.

Van Dang and W Bruce Croft. 2012. Diversity by proportionality: an election-based approach to search result diversification. In *Proc. SIGIR*. ACM, 65–74.

J Stephen Downie. 2008. The music information retrieval evaluation exchange (2005-2007): A window into music information retrieval research. *Acoustical Science and Technology* 29, 4 (2008), 247–255.

J Stephen Downie, Andreas F Ehmann, Mert Bay, and M Cameron Jones. 2010. The music information retrieval evaluation exchange: Some observations and insights. In *Advances in music information retrieval*. Springer, 93–115.

Hui Fang, Hao Wu, Peilin Yang, and ChengXiang Zhai. 2014. Virlab: A web-based virtual lab for learning and studying information retrieval models. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 1249–1250.

Nicola Ferro and Gianmaria Silvello. 2015. Rank-Biased Precision Reloaded: Reproducibility and Generalization. In *Advances in Information Retrieval*. Springer, 768–780.

Norbert Fuhr. 2012. Salton Award Lecture: Information Retrieval As Engineering Science. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 1–2. DOI:http://dx.doi.org/10.1145/2348283.2348285

Matthias Hagen, Martin Potthast, Michel Büchner, and Benno Stein. 2015. Twitter sentiment detection via ensemble classification using averaged confidence scores. In *Advances in Information Retrieval*. Springer, 741–754.

Jiyin He, Vera Hollink, and Arjen de Vries. 2012. Combining implicit and explicit topic representations for result diversification. In *Proc. SIGIR*. ACM, 851–860.

Xuedong Huang, James Baker, and Raj Reddy. 2014. A historical perspective of speech recognition. *Commun. ACM* 57, 1 (2014), 94–103.

Samuel Huston and W Bruce Croft. 2014. A Comparison of Retrieval Models using Term Dependencies. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 111–120.

Johannes Leveling, Lorraine Goeuriot, Liadh Kelly, and Gareth J Jones. 2012. DCU@ TRECMed 2012: Using ad-hoc baselines for domain-specific retrieval. In *Proceedings of TREC*.

Philipp Mayr, Andrea Scharnhorst, Birger Larsen, Philipp Schaer, and Peter Mutschke. 2014. *Bibliometric-enhanced information retrieval*. Springer, 798–801.

http://link.springer.com/chapter/10.1007/978-3-319-06028-6_99
Donald Metzler and Oren Kurland. 2012. Experimental Methods for Information Retrieval. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 1185–1186. DOI:http://dx.doi.org/10.1145/2348283.2348534
David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 215–224.
Antti Puurula. 2013. Cumulative progress in language models for information retrieval. In *Proceedings of Australasian Language Technology Association Workshop*. 96–100.
Jinfeng Rao, Jimmy Lin, and Miles Efron. 2015. Reproducible Experiments on Lexical and Temporal Feedback for Tweet Search. In *Advances in Information Retrieval*. Springer, 755–767.
Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. *NIST SPECIAL PUBLICATION SP* (1995), 109–109.
Alan Said and Alejandro Bellogín. 2014. Comparative recommender system evaluation: benchmarking recommendation frameworks. In *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 129–136.
Tetsuya Sakai and Chin-Yew Lin. 2010. Ranking Retrieval Systems without Relevance Assessments-Revisited. In *The Third International Workshop on Evaluating Information Access (EVIA)*. 25–33.
Mark Sanderson. 2010. Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends in Information Retrieval* 4, 4 (2010), 247375.
Mark Sanderson, Andrew Turpin, Ying Zhang, and Falk Scholer. 2012. Differences in effectiveness across sub-collections. In *Proc. CIKM*. ACM, 1965–1969.
S. Sanner, S. Guo, T. Graepel, S. Kharazmi, and S. Karimi. 2011. Diverse Retrieval via Greedy Optimization of Expected 1-call@ k in a Latent Subtopic Relevance Model. In *Proc. CIKM*.
Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2011a. How diverse are web search results?. In *Proc. SIGIR*. ACM, 1187–1188.
Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2011b. Intent-aware search result diversification. In *Proc. SIGIR*. ACM, 595–604.
Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2011c. On the suitability of diversity metrics for learning-to-rank for diversity. In *Proc. SIGIR*. ACM, 1185–1186.
Rodrygo LT Santos, Craig Macdonald, and Iadh Ounis. 2012. On the role of novelty for search result diversification. *Information retrieval* 15, 5 (2012), 478–502.
Rodrygo L. T. Santos, Jie Peng, Craig Macdonald, and Iadh Ounis. 2010. Explicit search result diversification through sub-queries. In *Proc. ECIR*. Springer, Milton Keynes, UK, 87–99.
Markus Schedl, Emilia Gómez, and Julián Urbano. 2014. Music Information Retrieval: Recent Developments and Applications. *Foundations and Trends in Information Retrieval* (2014).
Aliaksei Severyn, Alessandro Moschitti, Manos Tsagkias, Richard Berendsen, and Maarten de Rijke. 2014. A Syntax-Aware Re-ranker for Microblog Retrieval. In *Proc. SIGIR*. ACM.
K. Spärck Jones and C. J. van Rijsbergen. 1976. Information Retrieval Test Collections. *Journal of Documentation* 32 (1976), 59 75.
Florian Stegmaier, Christin Seifert, Roman Kern, Patrick Hfler, Sebastian Bayerl, Michael Granitzer, Harald Kosch, Stefanie Lindstaedt, Belgin Mutlu, Vedran Sabol, and others. 2014. *Unleashing semantics of research data*. Springer, 103–112.

http://link.springer.com/chapter/10.1007/978-3-642-53974-9_10
Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. 2005. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, Vol. 2. Citeseer, 26. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.65.3502&rep=rep1&type=pdf
Andrew Trotman and David Keeler. 2011. Ad hoc ir: not much room for improvement. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 10951096. http://dl.acm.org/citation.cfm?id=2010066
Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and Language Models Examined. In *Proceedings of the 2014 Australasian Document Computing Symposium (ADCS '14)*. ACM, 58:58–58:65. DOI:http://dx.doi.org/10.1145/2682862.2682863
David Vallet and Pablo Castells. 2012. Personalized diversification of search results. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 841–850.
Saúl Vargas, Pablo Castells, and David Vallet. 2012. Explicit relevance models in intent-oriented information retrieval diversification. In *Proc. SIGIR*. ACM, 75–84.
Ellen M Voorhees and Chris Buckley. 2002. The effect of topic set size on retrieval experiment error. In *Proc. SIGIR*. ACM, 316–323.
E. M. Voorhees and D. K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval* (illustrated edition ed.). The MIT Press.
J. Wang and J. Zhu. 2009. Portfolio theory of information retrieval. In *Proc. SIGIR*. ACM, 115–122.
William Webber, Alistair Moffat, and Justin Zobel. 2008. Score standardization for inter-collection comparison of retrieval systems. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 51–58.
Zheng Ye, Jimmy Xiangji Huang, and Jun Miao. 2012. A hybrid model for ad-hoc information retrieval. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1025–1026.
C.X. Zhai and J. Lafferty. 2006. A risk minimization framework for information retrieval. *Information Processing & Management* 42, 1 (2006), 31–55.
Guido Zuccon, Leif Azzopardi, Dell Zhang, and Jun Wang. 2012. Top-k retrieval using facility location analysis. In *Proc. ECIR*. Springer, 305–316.