TREC 2014 Web Track Overview

Kevyn Collins-Thompson
University of Michigan

Craig Macdonald
University of Glasgow

Paul Bennett
Microsoft Research

Fernando Diaz
Microsoft Research

Ellen M. Voorhees
NIST

February 18, 2015

1 Introduction

The goal of the TREC Web track over the past few years has been to explore and evaluate innovative retrieval approaches over large-scale subsets of the Web – currently using ClueWeb12, on the order of one billion pages. For TREC 2014, the sixth year of the Web track, we implemented the following significant updates compared to 2013. First, the *risk-sensitive retrieval* task was modified to assess the ability of systems to adaptively perform risk-sensitive retrieval against multiple baselines, including an optional self-provided baseline. In general, the risk-sensitive task explores the tradeoffs that systems can achieve between effectiveness (overall gains across queries) and robustness (minimizing the probability of significant failure, relative to a particular provided baseline). Second, we added query performance prediction as an optional aspect of the risk-sensitive task. The Adhoc task continued as for TREC 2013, evaluated using both adhoc and diversity relevance criteria.

This year, experiments by participating groups again used the ClueWeb12 Web collection, a successor to the ClueWeb09 dataset that comprises about one billion Web pages crawled between Feb-May 2012.[1] The crawling and collection process for ClueWeb12 included a rich set of seed URLs based on commercial search traffic, Twitter and other sources, and multiple measures for flagging undesirable content such as spam, pornography, and malware.

For consistency with last year's Web track, topic development was done using a very similar process to the one used in 2013. A common topic set

[1] Details on ClueWeb12 are available at `http://boston.lti.cs.cmu.edu/clueweb12`

of 50 additional new topics was developed and used for both the Adhoc and Risk-sensitive tasks. In keeping with the goal of reflecting authentic Web retrieval problems, the Web track topics were again developed from a pool of candidate topics based on the logs and data resources of commercial search engines. The initial set of candidates developed for the 2013 track was large enough that candidate topics not used in 2013 were used as the pool for the 2014 track. We kept the distinction between faceted topics, and unfaceted (single-facet) topics. Faceted topics were more like "head" queries, and structured as having a representative set of subtopics, with each subtopic corresponding to a popular subintent of the main topic. The faceted topic queries had subintents that were likely to be most relevant to users. Unfaceted (single-facet) topics were intended to be more like "tail" queries with a clear question or intent. For faceted topics, query clusters were developed and used by NIST for topic development. Only the base query was released to participants initially: the topic structures containing subtopics and single- vs multi-faceted *vs.* topic type were only released after runs were submitted. This was done to avoid biases that might be caused by revealing extra information about the information need that may not be available to Web search systems as part of the actual retrieval process.

The Adhoc task judged documents with respect to the topic as a whole. Relevance levels are similar to the levels used in commercial Web search, including a spam/junk level. The top two levels of the assessment structure are related to the older Web track tasks of homepage finding and topic distillation. Subtopic assessment was also performed for the faceted topics, as described further in Section 3.

Table 1 summarizes participation in the TREC 2014 Web Track. Overall, we received 42 runs from 9 groups: 30 ad hoc runs and 12 risk-sensitive runs. The number of participants in the Web track decreased over 2013 (when 15 groups participated, submitting 61 runs). Seven runs were categorized as manual runs (4 adhoc, 3 risk), submitted across 2 groups: all other runs were automatic with no human intervention. All submitted runs used the main Category A corpus: none used the Category B subset of ClueWeb12.

The submitting groups were:

Task	Adhoc	Risk	Total
Groups	9	5	9
Runs	30	12	42

Table 1: TREC 2014 Web Track participation.

Carnegie Mellon University and Ohio State University
Chinese Academy of Sciences
Delft University of Technology
Medical Informatics Laboratory
University of Delaware (Carterette)
University of Delaware (Fang)
University of Glasgow (Terrier Team)
University of Massachusetts Amherst
University of Twente

Three teams submitted at least one run with an associated Query Performance Prediction file.

In the following, we recap on the corpus (Section 2), and topics (Section 3) used for TREC 2014. Section 4 details the pooling and evaluation methodologies applied for Adhoc and Risk-Sensitive tasks, as well as the results of the participating groups. Section 5 examines sources of variation across submitted runs using Principal Components Analysis. Section 6 details the efforts of participants on the query performance sub-task. Concluding remarks follow in Section 7.

2 ClueWeb12 Category A and B corpus

As with ClueWeb09, the ClueWeb12 corpus comes with two datasets: Category A, and Category B. The Category A dataset is the main corpus and contains about 733 million documents (27.3 TB uncompressed, 5.54 TB compressed). The Category B dataset is a sample from Category A, containing about 52 million documents, or about 7% of the Category A total. Details on how the Category A and B corpora were created may be found on the Lemur project website[2]. We strongly encouraged participants to use the full Category A data set if possible. All of the results in this overview paper are labeled by their corpus category.

[2] `http://lemurproject.org/clueweb12/specs.php`

3 Topics

NIST created and assessed 50 new topics for the TREC 2014 Web track. As with TREC 2013, the TREC 2014 Web track included a significant proportion of more focused topics designed to represent more specific, less frequent, possibly more difficult queries. To retain the Web flavor of queries in this track, we kept the notion that some topics may be multi-faceted, i.e. broader in intent and thus structured as a representative set of subtopics, each related to a different potential aspect of user need. Examples are provided below. For topics with multiple subtopics, documents were judged with respect to each of the subtopics. For each subtopic, NIST assessors made a six-point judgment scale as to whether or not the document satisfied the information need associated with the subtopic. For those topics with multiple subtopics, the set of subtopics was intended to be representative, not exhaustive.

Subtopics were based on information extracted from the logs of a commercial search engine, based on a pool of remaining topic candidates created but not sampled for the 2013 Web track. Topics having multiple subtopics had subtopics selected roughly by overall popularity, which was achieved using combined query suggestion and completion data from two commercial search engines. In this way, the focus was retained on a balanced set of popular subtopics, while limiting the occurrence of strange and unusual interpretations of subtopic aspects. Single-facet topic candidates were developed based on queries extracted from search log data that were low-frequency ('tail-like') but issued by multiple users; less than 10 terms in length; and relatively low effectiveness scores across multiple commercial search engines (as of January 2013).

The topic structure was similar to that used for the TREC 2009 topics. An example of a single-facet topic:

```
<topic number="293" type="single">
<query>educational advantages of social networking sites</query>
<description>
What are the educational benefits of social networking sites?
</description>
</topic>
```

An example of a faceted topic:

```
<topic number="289" type="faceted">
```

```
<query>benefits of yoga</query>
<description>What are the benefits of yoga for kids?</description>
<subtopic number="1" type="inf">What are the benefits of yoga for kids?</subtopic>
<subtopic number="2" type="inf">Find information on yoga for seniors.</subtopic>
<subtopic number="3" type="inf">Does yoga help with weight loss?</subtopic>
<subtopic number="4" type="inf">What are the benefits of various yoga poses?</subtopic>
<subtopic number="5" type="inf">What are the benefits of yoga during pregnancy?</subtopic>
<subtopic number="6" type="inf">How does yoga benefit runners?</subtopic>
<subtopic number="7" type="inf">Find the benefits of yoga nidra.</subtopic>
</topic>
```

The initial release of topics to participants included only the query field, as shown in the excerpt here:

```
251:identifying spider bites
252:history of orcas island
253:tooth abscess
254:barrett's esophagus
255:teddy bears
```

As shown in the above examples, those topics with a clear focused intent have a single subtopic. Topics with multiple subtopics reflect underspecified queries, with different aspects covered by the subtopics. We assume that a user interested in one aspect may still be interested in others. Each subtopic was informally categorized by NIST as being either navigational ("nav") or informational ("inf"). A navigational subtopic usually has only a small number of relevant pages (often one). For these subtopics, we assume the user is seeking a page with a specific URL, such as an organization's homepage. On the other hand, an informational query may have a large number of relevant pages. For these subtopics, we assume the user is seeking information without regard to its source, provided that the source is reliable.

For the adhoc task, relevance is judged on the basis of the description field. Thus, the first subtopic is always identical to the description sentence.

4 Methodology and Measures

4.1 Pooling and Judging

For each topic, participants in the adhoc and risk-sensitive tasks submitted a ranking of the top 10,000 results for that topic. All submitted runs were included in the pool for judging (with the exception of 2 runs from 1 group that were marked as lowest judging priority and exceeded the per-team task limit in the guidelines). A common pool was created from the runs submitted to both tasks, which were pooled to rank depth 25.

For the risk-sensitive task, versions of `ndeval` and `gdeval` supporting the risk-sensitive versions of the evaluation measures (described below) were provided to NIST. These versions were identical to those used in last year's track except for a minor adjustment in output formatting.

All data and tools required for evaluation, including the scoring programs ndeval and gdeval as well as the baseline runs used in computation of the risk-sensitive scores are available in the track's github distribution[3].

The relevance judgment for a page was one of a range of values as described in Section 4.2. All topic-aspect combinations this year had a non-zero number of known relevant documents in the ClueWeb12 corpus. For topics that had a single aspect in the original topics file, that one aspect is used. For all other topics, aspect number 1 is the single aspect. All topics were judged to depth 25.

4.2 Adhoc Retrieval Task

An adhoc task in TREC provides the basis for evaluating systems that search a static set of documents using previously-unseen topics. The goal of an adhoc task is to return a ranking of the documents in the collection in order of decreasing probability of relevance. The probability of relevance for a document is considered independently of other documents that appear before it in the result list. For the adhoc task, documents are judged on the basis of the description field using a six-point scale, defined as follows:

1. Nav: This page represents a home page of an entity directly named by the query; the user may be searching for this specific page or site. (relevance grade 4)

2. Key: This page or site is dedicated to the topic; authoritative and comprehensive, it is worthy of being a top result in a web search engine.

[3] `http://github.com/trec-web/trec-web-2014`

(relevance grade 3)

3. HRel: The content of this page provides substantial information on the topic. (relevance grade 2)

4. Rel: The content of this page provides some information on the topic, which may be minimal; the relevant information must be on that page, not just promising-looking anchor text pointing to a possibly useful page. (relevance grade 1)

5. Non: The content of this page does not provide useful information on the topic, but may provide useful information on other topics, including other interpretations of the same query. (relevance grade 0)

6. Junk: This page does not appear to be useful for any reasonable purpose; it may be spam or junk (relevance grade -2).

After each description we list the relevance grade assigned to that level as they appear in the judgment (qrels) file. These relevance grades are also used for calculating graded effectiveness measures, except that a value of -2 is treated as 0 for this purpose.

The primary effectiveness measure for the adhoc task was *intent-aware expected reciprocal rank* (ERR-IA) which is a diversity-based variant of ERR as defined by Chapelle et al. [1] that accounts for faceted topics. For single-facet topics, ERR-IA simply becomes ERR. We also report an intent-aware version of nDCG, α-nDCG [3], and novelty- and rank-biased precision (NRBP) [2]. Table 2 presents the (diversity-aware) performance of the participating groups in the Adhoc task, ranked by ERR-IA@20 and selecting each group's highest performing run among those they submitted to the Adhoc task. The applied measures, ERR-IA@20, α-nDCG@20, and NRBP, take into account the multiple possible subintents underlying a given topic, and hence measure if the participants systems would have performed effective retrieval for such multi-faceted queries. Of note, the highest performing run was a manual run. Moreover, while category B runs were permitted, no participanting groups chose to submit category B runs.

We also report the standard (non-diversity-based) ERR@20 and nDCG@20 effectiveness measures for the Adhoc task in Table 3. We note that these rankings exhibit some differences from Table 2, demonstrating that some systems may focus upon single dominant interpretations of a query, without trying to uncover other possible interpretations.

Finally, Figure 1 visualizes the per-topic variability in ERR@20 across all submitted runs. For many topics, there was relatively little difference

Table 2: Top ad-hoc task results (diversity-based measures), ordered by ERR-IA@20. Only the best automatic run according to ERR-IA@20 from each group is included in the ranking. Only one team submitted a manual run that outperformed automatic – the highest manual run from that team (udel_fang) is included as well.

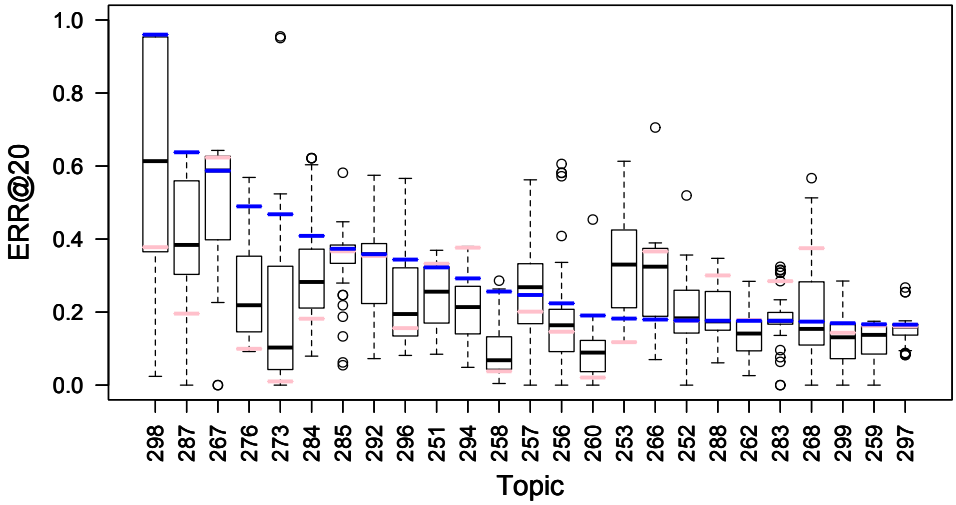
Group	Run	Cat	Type	ERR-IA@20	α-nDCG@20	NRBP
udel_fang	UDInfoWebLES	A	manual	0.688	0.754	0.656
udel_fang	UDInfoWebAX	A	auto	0.608	0.694	0.564
uogTr	uogTrDwl	A	auto	0.595	0.682	0.548
BUW	webisWt14axMax	A	auto	0.589	0.667	0.550
udel	udelCombCAT2	A	auto	0.583	0.656	0.545
wistud	wistud.runB	A	auto	0.583	0.660	0.543
ICTNET	ICTNET14ADR3	A	auto	0.580	0.652	0.541
Group.Xu	Terran	A	auto	0.578	0.647	0.541
UMASS_CIIR	CiirAll1	A	auto	0.558	0.639	0.512
Organizers1	TerrierBase	A	auto	0.542	0.627	0.501
ut	utexact	A	auto	0.535	0.612	0.494
SNUMedinfo	SNUMedinfo12	A	auto	0.531	0.624	0.481
Organizers2	IndriBase	A	auto	0.513	0.585	0.474

between the top runs and the median, according to some of the effectiveness measures (e.g. ERR@20 and some diversity measures). As a result, a small number of topics tended to contribute to most of the variability observed between systems. In particular, topics 298, 273, 253, 293, 269 had especially high variability across systems. In comparing the Indri and Terrier baselines used for risk-sensitive evaluation: absolute difference in ERR@20 between the baselines was greater than 0.10 for 17 topics, and greater than 0.20 for 7 topics. Expressed as a relative percentage gain/loss, there were 18 topics for which the Terrier baseline ERR@20 was at least 50% higher than the Indri baseline, and 6 topics where the Indri baseline was at least 50% higher than the Terrier baseline.

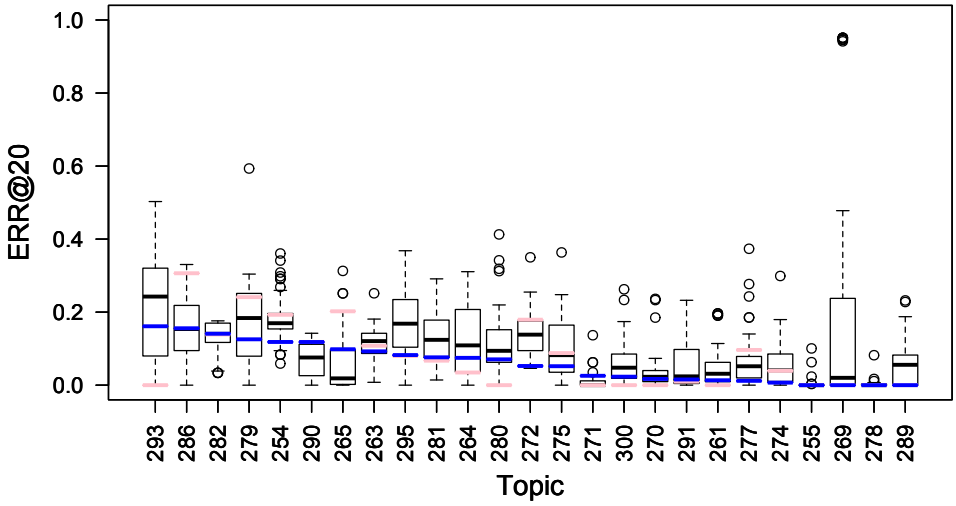
4.3 Risk-sensitive Retrieval Task

The *risk-sensitive retrieval* task for Web evaluation rewards algorithms that not only achieve improvements in average effectiveness across topics (as in the adhoc task), but also maintain good robustness, which we define as *minimizing the risk of significant failure* relative to a given baseline.

Search engines use increasingly sophisticated stages of retrieval in their quest to improve result quality: from personalized and contextual re-ranking to automatic query reformulation. These algorithms aim to increase retrieval



(a) Top 25 topics (by descending Terrier baseline ERR@20)



(b) Bottom 25 topics (by descending Terrier baseline ERR@20)

Figure 1: Boxplots for TREC 2014 Web topics, showing variation in ERR@20 effectiveness across all submitted runs. Topics are sorted by decreasing Terrier baseline ERR@20 (blue bar) with Indri baseline also shown (light pink bar).

Table 3: Top ad-hoc task results (non-diversity-based) ordered by ERR@20. Only the best automatic run according to ERR@20 from each group is included in the ranking. Only one team submitted a manual run that outperformed automatic – the highest manual run from that team (udel_fang) is included as well.

Group	Run	Cat	Type	ERR@20	nDCG@20
udel_fang	UDInfoWebRiskTR	A	manual	0.233	0.325
ICTNET	ICTNET14ADR1	A	auto	0.208	0.261
udel_fang	UDInfoWebAX	A	auto	0.207	0.307
Group.Xu	Terran	A	auto	0.204	0.294
uogTr	uogTrDwl	A	auto	0.195	0.324
Organizers1	TerrierBase	A	auto	0.189	0.260
udel	udelCombCAT2	A	auto	0.179	0.261
SNUMedinfo	SNUMedinfo12	A	auto	0.176	0.270
BUW	webisWt14axAll	A	auto	0.174	0.258
wistud	wistud.runB	A	auto	0.174	0.291
ut	utexact	A	auto	0.172	0.226
Organizers2	IndriBase	A	auto	0.153	0.243
UMASS_CIIR	CiirAll1	A	auto	0.153	0.250

effectiveness on average across queries, compared to a baseline ranking that does not use such operations. However, these operations are also risky since they carry the possibility of failure – that is, making the results worse than if they had not been used at all. The goal of the risk-sensitive task is two-fold: 1) To encourage research on algorithms that go beyond just optimizing average effectiveness in order to effectively optimize both effectiveness and robustness, and achieve effective tradeoffs between these two competing goals; and 2) to explore effective risk-aware evaluation criteria for such systems.

The risk-sensitive retrieval track is related to the goals of the earlier TREC Robust Track (TREC 2004, 2005),[4] which focused on increasing retrieval effectiveness for poorly-performing topics using evaluation measures such as geometric MAP that focused on maximizing the average improvement on the most difficult topics. The risk-sensitive retrieval track can be thought of as a next step in exploring more general retrieval objectives and evaluation measures that (a) explicitly account for, and can differentiate systems based on, differences in *variance* or other risk-related statistics of the win/loss distribution across topics for a single run, (b) the quality of *the*

[4] `http://trec.nist.gov/data/robust/04.guidelines.html`

curve derived from a set of tradeoffs between effectiveness and robustness achievable by systems, measured across *multiple runs* at different average effectiveness levels, and (c) computing (a) and (b) by accounting for the effectiveness of a competing baseline (both standard, and participant-supplied) as a factor in optimizing retrieval performance.

Two runs were provided as standard baselines, derived from the Indri[5] and Terrier[6] retrieval engines. For Indri, we used a pseudo-relevance feedback run as implemented by the Indri retrieval engine. Specifically, for each query, we used 10 feedback documents, 20 feedback terms, and a linear interpolation weight of 0.60 with the original query. For the Terrier standard baseline, documents were ranked using the DPH weighting model from the Divergence from Randomness framework. For both systems, we provided baselines with and without application of the Waterloo spam classifier, where the filtered runs removed all documents with a percentile-score less than 70 from the rankings[7].

As with the adhoc task, we use Intent-Aware Expected Reciprocal Rank (ERR-IA) as the basic measure of retrieval effectiveness, and per-query retrieval delta is defined as the absolute difference in effectiveness between a contributed run and the above standard baseline run, for a given query. A positive delta means a win for the system on that query, and negative delta means a loss. For single runs, the following will be the main risk-sensitive evaluation measure. Let $\Delta(q) = R_A(q) - R_{BASE}(q)$ be the absolute win or loss for query q with system retrieval effectiveness $R_A(q)$ relative to the baseline's effectiveness $R_{BASE}(q)$ for the same query. We categorize the outcome for each query q in the set Q of all N queries according to the sign of $\Delta(q)$, giving three categories:
Hurt Queries (Q_-) have $\Delta(q) < 0$; Unchanged Queries (Q_0) have $\Delta(q) = 0$; Improved Queries (Q_+) have $\Delta(q) > 0$.
The risk-sensitive utility measure $U_{RISK}(Q)$ of a system over the set of queries Q is defined as:

$$U_{RISK}(Q) = 1/N \cdot [\Sigma_{q \in Q_+} \Delta(q) - (\alpha + 1)\Sigma_{q \in Q_-} \Delta(q)] \quad (1)$$

where α is the risk-aversion parameter. In words, this rewards systems that maximize average effectiveness, but also penalizes losses relative to the baseline results for the same query, weighting losses $\alpha + 1$ times as heavily as successes. For example, when the risk aversion parameter α is large, this

[5] `http://www.lemurproject.org/indri/`
[6] `http://terrier.org`
[7] `http://www.mansci.uwaterloo.ca/~msmucker/cw12spam/`

rewards systems that are more conservative and able to avoid large losses relative to the baseline. The adhoc task objective, maximizing only average effectiveness across queries, corresponds to the special case $\alpha = 0$.

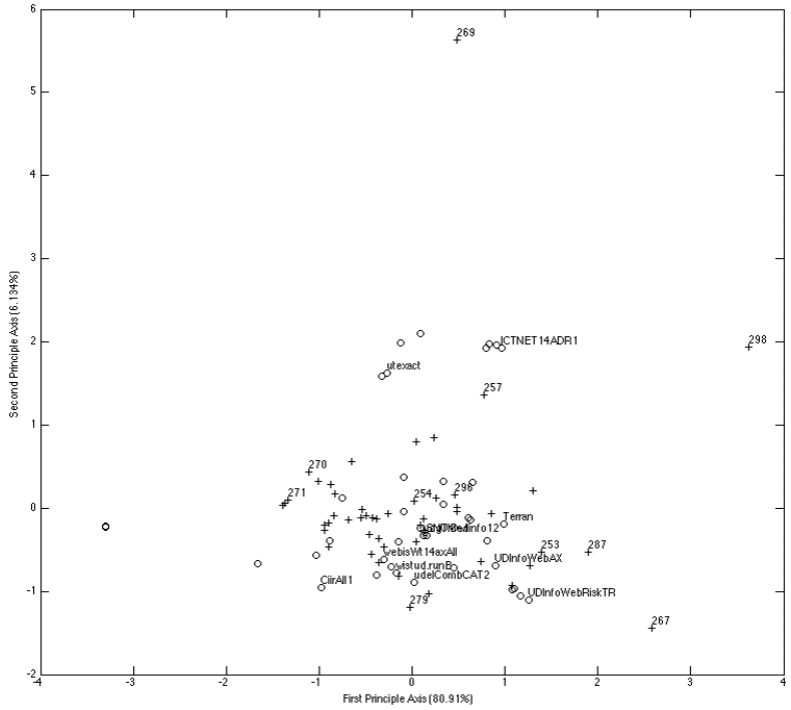
Table 5 summarizes the results for all runs submitted for the risk-sensitive retrieval task, according to the U_{RISK} evaluation measure with $\alpha = 5$, using each of the Indri and Terrier baselines. The average U_{RISK} across both baselines is also reported. Notably, the best average and Terrier-relative U_{RISK} was achieved by manual runs (in the top 3 places). However, the best Indri-relative U_{RISK} was achieved by an automatic run. Also, the relative ranking of runs by U_{RISK} changes considerably with the choice of baseline. This may indicate that some systems were tuned to optimize U_{RISK} for one baseline but not the other – or for no baseline at all.

In last year's risk-sensitive task, participants were asked to submit a set of runs that were optimized for different levels of risk aversion, e.g. by optimizing for U_{RISK} using different pre-specified values of α. However, not enough groups attempted this to allow for meaningful analysis. As a result, we adjusted the task this year so that participants were requested to submit risk-sensitive runs each of which was optimized against one of two different baselines. Participants also had the opportunity to include their own self-defined baseline runs. The goal was to reward systems that could successfully adapt to multiple baselines. Table 4 gives the complete set of U_{RISK} results ($\alpha = 5$) for the official and self-defined baselines.

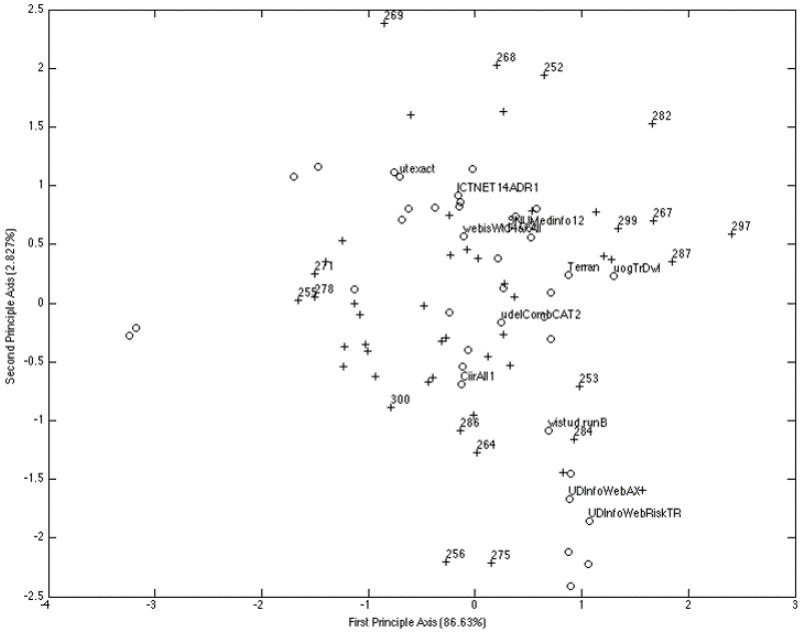
5 Analysis of variation in effectiveness across runs

Using Principal Component Analysis (PCA) on topic retrieval effectiveness scores across all submitted runs can help identify underlying factors that account for variation across systems. Figure 2 shows a biplot based on PCA of standardized retrieval effectiveness scores (ERR@20 and nDCG@20) for all 2014 topics and runs. Runs are plotted as circles, with selected runs labeled with TREC run name. Topics are plotted as plus signs. (Details on PCA plots for IR can be found in Dinçer [5].)

Topics and runs near the origin have a mean effectiveness score close to the group mean. Topics and runs far from the origin along one or both component dimensions are those that have more influence on the group ranking (based on average ERR@20 or nDCG@20 respectively). Topics that are close together in the plot are those with similar effectiveness profiles across runs. Likewise, runs that are close together had similar effectiveness profiles across topics.



(a) Measure: ERR@20



(b) Measure: NDCG@20

Figure 2: Principal components biplot of TREC 2014 Web Track runs and topics, based on ERR@20 (top) and nDCG@20 (bottom).

More than 80% of the overall variation in retrieval effectiveness was captured by a single (first) principal component, for both ERR@20 (80.9%) and nDCG@20 (86.6%). A subset of about 10 topics had large coordinate values (greater than 1) associated with this first principal component, and thus had a large effect on the overall ranking of systems by average retrieval effectiveness. Examples of such topics include topic 267 (feliz navidad lyrics) for ERR@20 and topic 297 (altitude sickness) for NDCG@20. (These results are in accord with those of Figure 1 that visualizes per-topic variability across runs.)

The second principal component accounted for much less remaining variation for both ERR@20 (6.1%) and nDCG@20 (2.8%). A set of 3-5 topics (ERR@20 vs nDCG@20) were strongly associated with the second component. These included topic 269 (marshall county schools) and topic 298 (medical care and jehovah's witnesses). We can see from the biplot that some runs (e.g. UDInfoWebRiskTR) did well on the first-component topics (and these were typically highly-ranked overall) but less so on the second-component topics, while other runs (e.g. utexact and ICTNET14ADR1) did well on the second-component topics.

The separation between first- and second-component runs could be due to different choices of text representation or feature weighting. For example, similarities between the runs that did well on second-component topics (utexact and ICTNET14ADR1) might be due to their focus on anchor text. Better understanding of these factors is a topic for future investigation.

6 Query Performance Prediction

Determining the effectiveness of a retrieval, be it a baseline or a treatment, is one of the fundamental subtasks in risk-aware ranking. As such, in addition to the core document ranking tasks, the web track included a query performance prediction subtask, where participants were asked to rank topics according to the predicted performance of the baseline and the treatment.

The format of the evaluation followed the TREC 2004 Robust Retrieval Track with minor variations [7]. Participants were asked to output a score for each topic: (a) a prediction score for the absolute effectiveness of the results for baseline used for risk-sensitive run, (b) a prediction score for the absolute effectiveness of the results for the risk-sensitive run, and (c) a relative gain or loss prediction score (the difference in effectiveness between the risk-sensitive run and the baseline run). We evaluated the ability of participants to predict the rank ordering of per-topic performance and, therefore, measured the

Kendall's τ correlation between the system performance prediction scores and the actual topic performance in terms of ERR@20.

We present summary results in Table 6 and scatterplots of per-topic prediction scores versus ERR@20 in Figures 3, 4, and 5.

We found that participant performance predictors were weakly correlated with the actual topic performance. On an absolute scale, the values of Kendall's τ are low, with an mean τ of < 0.1 across all prediction tasks. The Kendall's τ values tend to lie between 0.1 and 0.5 for web topics [6, Table 4] and 0.3 and 0.5 for non-web tasks [4, Table 1], suggesting that the 2014 topics may be more confusing to systems than previous collections.

For both baseline and risk runs in isolation, participant predictions were more weakly correlated ($\overline{\tau}_{\text{baseline}} = 0.04$, $\overline{\tau}_{\text{riskrun}} = 0.03$) compared to predictions of relative improvements ($\overline{\tau}_{\text{relative}} = 0.07$). This observation suggests that detecting within-topic performance may be easier and, as a result, supports risk-sensitive retrieval as a research direction.

7 Conclusions

This is the last year for the Web track at TREC in its current form. Over the past 6 years, the Web track has developed resources including 300 topics with associated relevance judgments, across two separate corpora: ClueWeb09 and ClueWeb12.

Particular areas that we believe that have benefited from the TREC Web track include approaches for learning-to-rank, diversification and the efficiency/effectiveness tradeoff, as well as risk-sensitive models and evaluation. In particular, we believe that there is much further scope for promising research in the area of risk-sensitive retrieval, building upon the resources created by the TREC Web track.

8 Acknowledgements

We thank Jamie Callan, David Pane and the Language Technologies Institute at Carnegie Mellon University for creating and distributing the ClueWeb12 collection. This track could not operate without this valuable resource. We also thank Nick Craswell and Charlie Clarke for their many valuable suggestions and feedback.

References

[1] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 621–630, New York, NY, USA, 2009. ACM.

[2] C. Clarke, M. Kolla, and O. Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In L. Azzopardi, G. Kazai, S. Robertson, S. Rger, M. Shokouhi, D. Song, and E. Yilmaz, editors, *Advances in Information Retrieval Theory*, volume 5766 of *Lecture Notes in Computer Science*, pages 188–199. Springer Berlin Heidelberg, 2009.

[3] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 659–666, New York, NY, USA, 2008. ACM.

[4] F. Diaz. Performance prediction using spatial autocorrelation. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 583–590, New York, NY, USA, 2007. ACM Press.

[5] B. T. Dinçer. Statistical principal components analysis for retrieval experiments. *Journal of the American Society for Information Science and Technology*, 58(4):560–574, 2007.

[6] C. Hauff, V. Murdock, and R. Baeza-Yates. Improved query difficulty prediction for the web. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 439–448, New York, NY, USA, 2008. ACM.

[7] E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proceedings of the Thirteenth Text REtrieval Conference, TREC 2004*, pages 70–79, 2005.

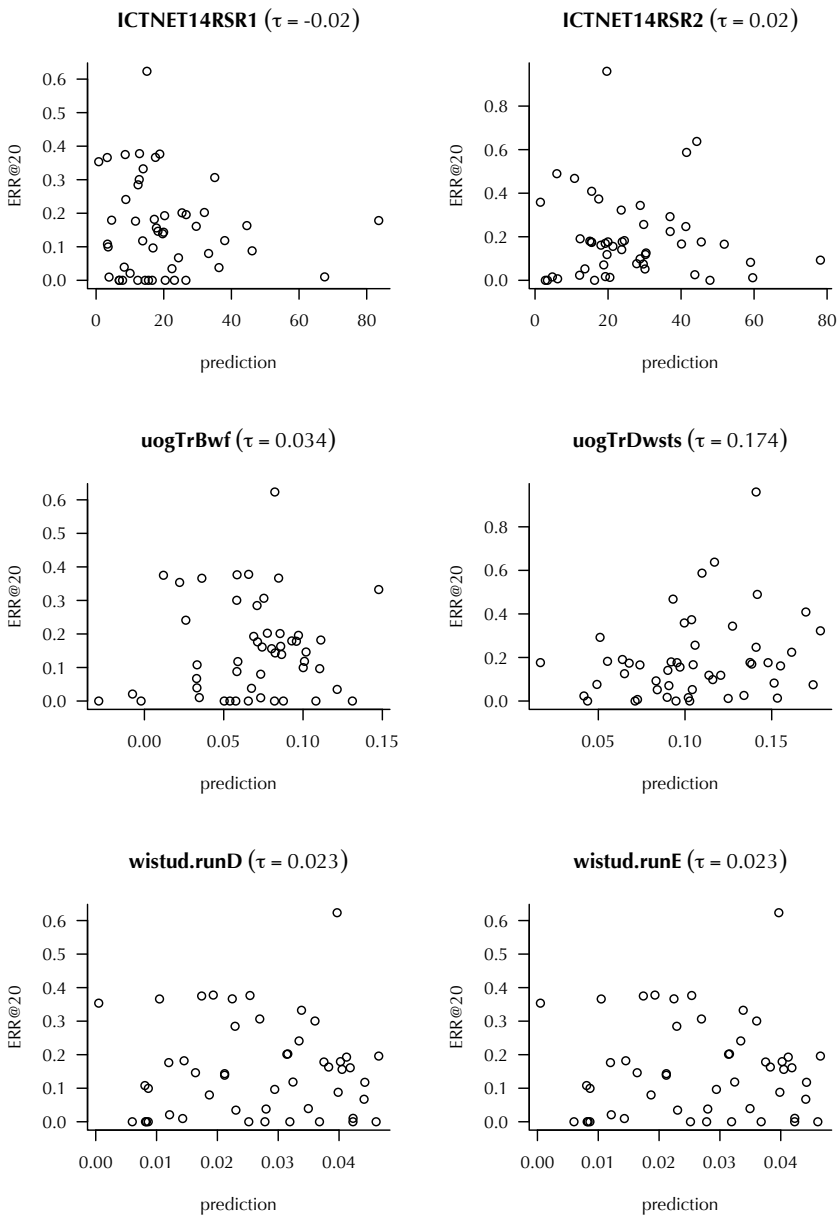


Figure 3: Baseline Performance Prediction Results

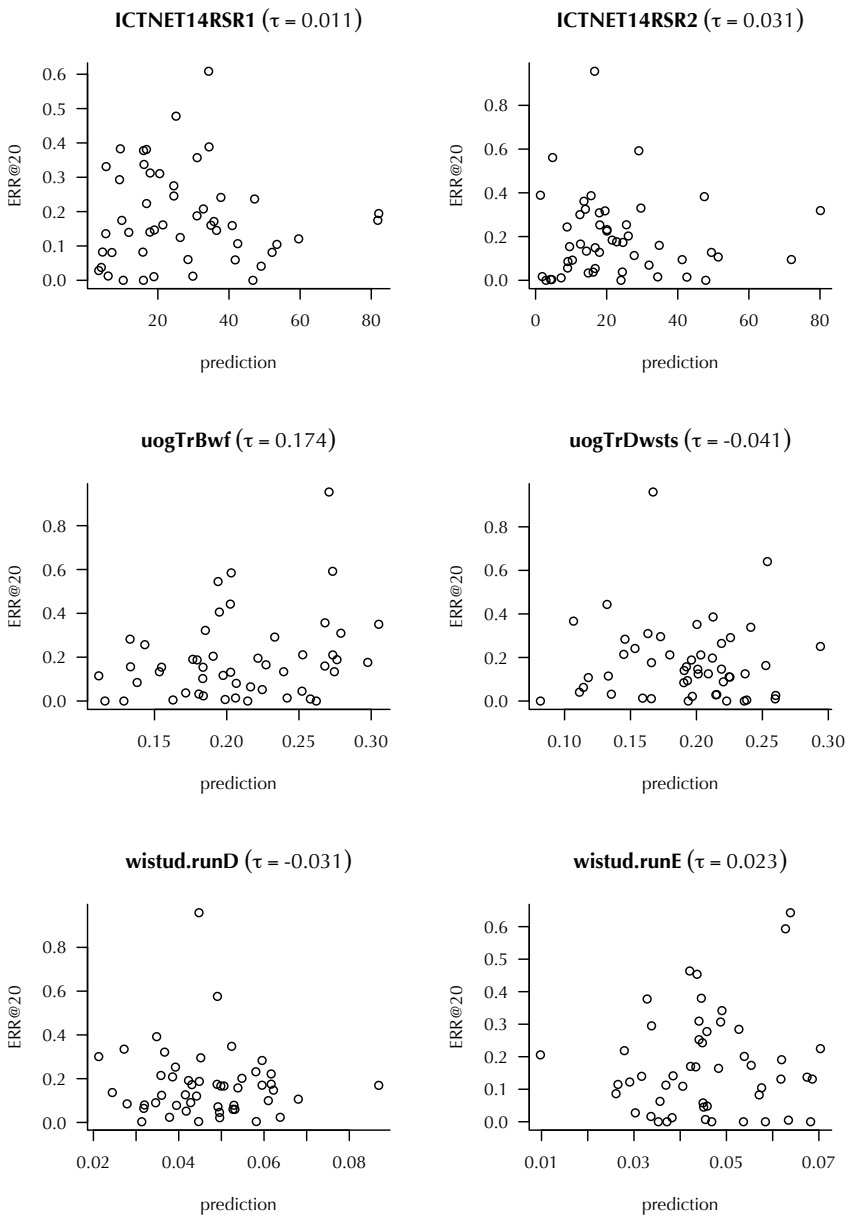


Figure 4: Risk Run Performance Prediction Results

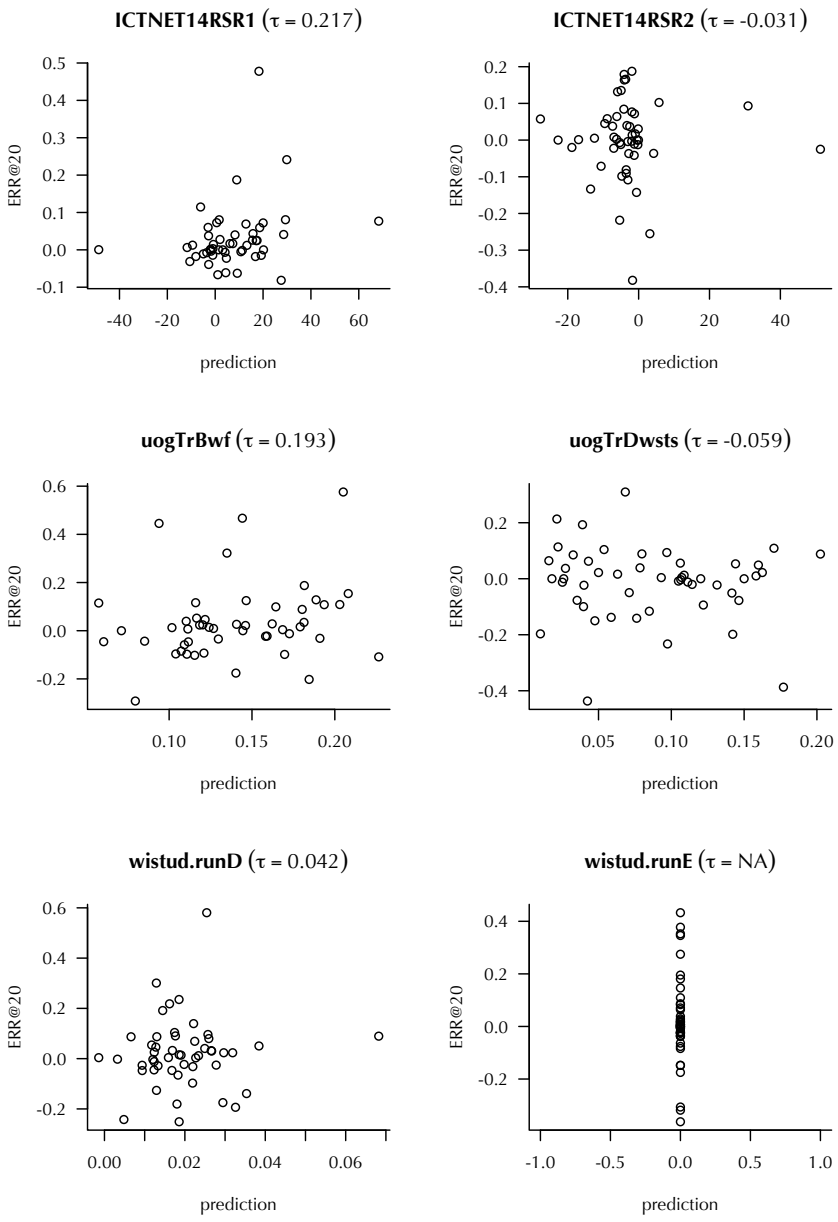


Figure 5: Relative Performance Prediction Results

Table 4: Risk-sensitive effectiveness of submitted risk runs relative to official and self-defined baselines.

Run	Baseline	U_{RISK}, $\alpha = 5$
ICTNET14RSR1	indri	-0.01702
ICTNET14RSR1	terrier	-0.25635
ICTNET14RSR1	ICTNET14ADR1	-0.32634
ICTNET14RSR1	ICTNET14ADR2	-0.31050
ICTNET14RSR1	ICTNET14ADR3	-0.32254
ICTNET14RSR2	indri	-0.11972
ICTNET14RSR2	terrier	-0.18145
ICTNET14RSR2	ICTNET14ADR1	-0.32083
ICTNET14RSR2	ICTNET14ADR2	-0.30409
ICTNET14RSR2	ICTNET14ADR3	-0.31601
ICTNET14RSR3	indri	-0.15928
ICTNET14RSR3	terrier	-0.20986
ICTNET14RSR3	ICTNET14ADR1	-0.02661
ICTNET14RSR3	ICTNET14ADR2	-0.01318
ICTNET14RSR3	ICTNET14ADR3	-0.01554
udelCombCAT2	indri	-0.11092
udelCombCAT2	terrier	-0.27713
udelCombCAT2	udel_itu	-0.37582
udelCombCAT2	udel_itub	-0.32152
UDInfoWebRiskAX	indri	-0.09497
UDInfoWebRiskAX	terrier	-0.13282
UDInfoWebRiskAX	UDInfoWebAX	-0.09389
UDInfoWebRiskAX	UDInfoWebENT	-0.09006
UDInfoWebRiskAX	UDInfoWebLES	-0.06893
UDInfoWebRiskRM	indri	-0.07929
UDInfoWebRiskRM	terrier	-0.12730
UDInfoWebRiskRM	UDInfoWebAX	-0.09723
UDInfoWebRiskRM	UDInfoWebENT	-0.08083
UDInfoWebRiskRM	UDInfoWebLES	-0.07408
UDInfoWebRiskTR	indri	-0.05661
UDInfoWebRiskTR	terrier	-0.10552
UDInfoWebRiskTR	UDInfoWebAX	-0.08702
UDInfoWebRiskTR	UDInfoWebENT	-0.07484
UDInfoWebRiskTR	UDInfoWebLES	-0.06485
uogTrBwf	indri	-0.13225
uogTrBwf	terrier	-0.22992
uogTrBwf	uogTrDuax	-0.21253
uogTrBwf	uogTrDwl	-0.26402
uogTrBwf	uogTrIwa	-0.12952
uogTrDwsts	indri	-0.12092
uogTrDwsts	terrier	-0.26885
uogTrDwsts	uogTrDuax	-0.26911
uogTrDwsts	uogTrDwl	-0.27401
uogTrDwsts	uogTrIwa	-0.20727
uogTrq1	indri	-0.12489
uogTrq1	terrier	-0.22741
uogTrq1	uogTrDuax	-0.19293
uogTrq1	uogTrDwl	-0.22614
uogTrq1	uogTrIwa	-0.18079
wistud.runD	indri	-0.15582
wistud.runD	terrier	-0.23495
wistud.runD	wistud.runA	-0.04875
wistud.runD	wistud.runB	-0.18761
wistud.runD	wistud.runC	-0.18761
wistud.runE	indri	-0.17293
wistud.runE	terrier	-0.35354
wistud.runE	wistud.runA	-0.21114
wistud.runE	wistud.runB	-0.19061
wistud.runE	wistud.runC	-0.19061

Table 5: Risk-sensitive effectiveness of submitted risk runs, sorted by average descending U_{RISK} across baselines with $\alpha = 5$.

Run	Type	Indri U_{RISK}	Terrier U_{RISK}	Average U_{RISK}
UDInfoWebRiskTR	manual	-0.057	-0.106	-0.081
UDInfoWebRiskRM	manual	-0.079	-0.127	-0.103
UDInfoWebRiskAX	manual	-0.095	-0.133	-0.114
ICTNET14RSR1	automatic	-0.017	-0.256	-0.137
ICTNET14RSR2	automatic	-0.120	-0.182	-0.150
uogTrq1	automatic	-0.125	-0.227	-0.176
uogTrBwf	automatic	-0.132	-0.230	-0.181
ICTNET14RSR3	automatic	-0.159	-0.210	-0.185
udelCombCAT2	automatic	-0.111	-0.277	-0.194
uogTrDwsts	automatic	-0.121	-0.270	-0.195
wistud.runD	automatic	-0.156	-0.235	-0.195
wistud.runE	automatic	-0.173	-0.354	-0.263

	baseline	riskrun	relative
ICTNET14RSR1	-0.0200	0.0114	0.2169
ICTNET14RSR2	0.0196	0.0311	-0.0311
uogTrBwf	0.0341	0.1743	0.1927
uogTrDwsts	0.1744	-0.0409	-0.0588
wistud.runD	0.0233	-0.0310	0.0425
wistud.runE	0.0233	0.0230	-

Table 6: Kendall's τ between system predicted value of per-topic ERR@20 and observed ERR@20 for baseline and risk runs. The third column presents τ between the system predicted value of relative ERR@20 and observed relative ERR@20.