Overview of the TREC 2010 Web Track

Charles L. A. Clarke
University of Waterloo

Nick Craswell
Microsoft

Ian Soboroff
NIST

Gordon V. Cormack
University of Waterloo

1 Introduction

The TREC Web Track explores and evaluates Web retrieval technology over large collections of Web data. In its current incarnation, the Web Track has been active for two years. For TREC 2010, the track includes three tasks: 1) an adhoc retrieval task, 2) a diversity task, and 3) a spam task. As we did for TREC 2009, we based our experiments on the billion-page ClueWeb09[1] data set created by the Language Technologies Institute at Carnegie Mellon University.

The TREC 2009 Web Track included a traditional adhoc retrieval task, employing topical binary relevance assessments and reporting estimated MAP as its primary effectiveness measure [4]. For TREC 2010, we modified this traditional assessment process to incorporate multiple relevance levels, which are similar in structure to the levels used in commercial Web search. This new assessment structure includes a spam/junk level, which also assisted in the evaluation of the spam task. The top two levels of the assessment structure are closely related to the homepage finding and topic distillation tasks appearing in older Web Tracks.

The diversity task was introduced for TREC 2009 and continues in TREC 2010, essentially unchanged [4]. The goal of this diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. The adhoc and diversity tasks share topics, which were developed by NIST with the assistance of information extracted from the the logs of a commercial Web search engine [9]. Topic creation and judging attempts to reflect a mix of genuine user requirements for the topic.

An analysis of last year's results indicates that the presence of spam and other low-quality pages substantially influenced the overall results [7]. This year we provided a preliminary spam ranking of the pages in the corpus, as an aid to groups who wish to reduce the number of low-quality pages in their results. The associated spam task required groups to provide their own ranking of the corpus according to "spamminess".

Table 1 summarizes participation in the TREC 2010 Web Track. A total of 23 groups participated in the track, a slight decrease from last year, when 26 groups participated. Many of the groups participating the diversity task also participated in the adhoc task, but not vice versa. The spam task attracted only 3 participants, including one group that participated only in this task. Only one group, ICTNET, participated in all three tasks.

[1] `boston.lti.cs.cmu.edu/Data/clueweb09`.

Task	Adhoc	Diversity	Spam	Total
Groups	20	12	3	23
Runs	55	32	5	92

Table 1: Participation in the TREC 2010 Web track

2 Category A and B Collections

The billion-page ClueWeb09 collection was crawled from the general Web during January and February 2009, and consists of 25TB of uncompressed data (5TB compressed) in multiple languages. Since some participants were not able to work with the full collection, the track accepted runs based on the smaller "Category B" subset of the full "Category A" collection. This Category B data set comprises about 50 million English-language pages, including the entirety of the English-language Wikipedia. Nonetheless, we strongly encouraged participants to use the full Category A data set, if possible. Results reported in this paper are labeled by their collection category.

3 Topics

NIST created and assessed 50 new topics for the track, which were used by both the adhoc task and diversity tasks. figure 1 provides two examples.

Each topic contains a query field, a description field, and several subtopic fields. The query field is intended to represent the text a user might enter into a Web search engine, if they were seeking the information indicated by the description field or by any of the subtopics. For the adhoc task, relevance is judged on the basis of the description field. For the diversity task, relevance is judged separately with respect to each subtopic. Initially, only the query field was released to track participants. The full topics were not released until the participants had submitted their experimental runs.

Each topic is assigned one of two types. Topics with ambiguous queries, such as topic 72 in figure 1, have several unrelated interpretations. One of these interpretations is chosen for the description, while a wider range of interpretations appear in the subtopics. Topics with faceted queries, such as topic 73 in the figure, have one primary interpretation, usually reflected in the description field. For these queries, the subtopics address narrower aspects of the broader topic.

Each subtopic is assigned one of two types. Navigational subtopics (with type "nav") assume the user is seeking a specific page or site. Navigational subtopics typically have a single relevant page. Informational subtopics (with type "inf") assume the user is seeking information without regard to its source, provided that the source is reliable. Informational subtopics typically have a large number of relevant pages.

Subtopics were developed with the assistance of query-term clusters extracted from the logs of a commercial Web search engine [9]. Topic development followed essentially the same procedure as last year [4]. Subtopics were chosen to be roughly balanced in terms of popularity. Strange and unusual aspects and interpretations were avoided as much as possible.

All topics are expressed in English. Non-English documents are never considered relevant, even if the assessor understands the language of the document and the document would be relevant in that language.

```
<topic number="72" type="ambiguous">
    <query>the sun</query>
    <description>
        Find information about the Sun, the star in our Solar System.
    </description>
    <subtopic number="1" type="inf">
        Find information about the Sun, the star in our Solar System.
    </subtopic>
    <subtopic number="2" type="nav">
        Find the homepage for the U.K. newspaper, The Sun.
    </subtopic>
    <subtopic number="3" type="nav">
        Find the homepage for the Baltimore Sun newspaper.
    </subtopic>
</topic>

<topic number="73" type="faceted">
    <query>neil young</query>
    <description>
        Find music, tour dates, and information about the musician Neil Young.
    </description>
    <subtopic number="1" type="nav">
        Find albums by Neil Young to buy.
    </subtopic>
    <subtopic number="2" type="inf">
        Find biographical information about Neil Young.
    </subtopic>
    <subtopic number="3" type="nav">
        Find lyrics or sheet music for Neil Young's songs.
    </subtopic>
    <subtopic number="4" type="nav">
        Find a list of Neil Young tour dates.
    </subtopic>
</topic>
```

Figure 1: Examples of TREC 2010 Web track topics.

4 Tasks and Measures

4.1 Adhoc Task

An adhoc task in TREC investigates the performance of 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 of 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. (*grade 3*)

3. **HRel:** The content of this page provides substantial information on the topic. (*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. (*grade 1*)

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

6. **Junk:** This page does not appear to be useful for any reasonable purpose; it may be spam or junk. (*grade 0*)

After each description, we list the relevance grade assigned to that level for the purpose of calculating graded effectiveness measures.

The primary effectiveness measure for the adhoc task is *expected reciprocal rank* (ERR) as defined by Chapelle et al. [2]. We also report a variant of nDCG [8], as well as standard binary measures, including mean average precision (MAP) and precision at rank k (P@k). We compute ERR at rank k (ERR@k) as follows:

$$\text{ERR@}k \;=\; \sum_{i=1}^{k} \frac{R(g_i)}{i} \prod_{j=1}^{i-1} (1 - R(g_i)), \tag{1}$$

where $R(g) = \frac{2^g - 1}{16}$ and $g_1, g_2, ..., g_k$ are the relevance grades associated with the top k documents. We compute nDCG@k as $\frac{\text{DCG@}k}{\text{ideal DCG@}k}$, where

$$\text{DCG@}k \;=\; \sum_{i=1}^{k} \frac{2^{g_i} - 1}{\log_2(1 + i)}. \tag{2}$$

For the binary relevance measures, we treat grades 1-4 as relevant and grade 0 as non-relevant. We apply `trec_eval` to compute the binary measures.

4.2 Diversity Task

The diversity task is similar to the adhoc retrieval task, but differs in its judging criteria and evaluation measures. The goal of the diversity task is to return a ranked list of pages that together provide complete coverage for a query, while avoiding excessive redundancy in the result list. For this task, the probability of relevance of a document is conditioned on the documents that appear before it in the result list.

For the diversity task, documents are judged on the basis of the subtopics. For each subtopic, the assessor makes a binary judgment as to whether or not a document satisfies the information need associated with that subtopic.

The primary effectiveness measure for the adhoc task is a variant of *intent-aware expected reciprocal rank* (ERR-IA) as defined by Chapelle et al. [2]. We also report a number of other intent aware measures appearing in the literature, including α-nDCG@k [6], NRBP [5], and MAP-IA [1]. Clarke et al. [3] provide a detailed description and analysis of the novelty and diversity measures employed in the TREC Web track.

4.3 Spam Task

The goal of the spam task is to score each English-language document in the Category A ClueWeb09 collection according to how likely it is to be spam. For the purposes of this task, we employ a broad definition of spam, which comprises pages that are essentially junk, along with pages that are more obviously deceptive or detrimental. Spam had a major impact on the TREC 2009 adhoc submissions; practically every one was improved dramatically by the application of a spam filter [7].

Participant submissions were evaluated by how well they identified spam in the adhoc and diversity submissions, as measured by *area under the receiver operating characteristic curve* (AUC) [7]. For the purposes of the spam task evaluation, pages judged to be "Junk" under the six-point relevance scale described in Section 4.1 were considered to be spam, while all other judged pages were considered to be non-spam (i.e., "ham"). Unjudged pages were not considered in the computation of AUC.

5 Pooling and Judging

For each topic, participants in the adhoc and diversity tasks submitted a ranking of the top 10,000 documents for that topic. All submitted runs were included in the pool for judging. This year, a common pool was used for both tasks, and all runs were judged to depth 20 using both the adhoc and diversity judging criteria. In this paper, we report results only for runs explicitly submitted to one task or the other.

Group	Run	Cat	ERR@20	nDCG@20	P@20	MAP
msrsv	msrsv3	A	0.166	0.237	0.344	0.082
baseline	uwgym	A	0.164	0.241	0.374	0.069
umass	umassSDMW	A	0.138	0.293	0.484	0.148
isi	IvoryL2Rb	B	0.134	0.225	0.379	0.133
THUIR	THUIR10QaHt	A	0.128	0.201	0.331	0.112
uogTr	uogTrA42	A	0.127	0.245	0.411	0.127
IRRA	irra10b	B	0.126	0.260	0.443	0.133
unimelb	UMa10IASF	A	0.119	0.181	0.293	0.080
CMU_LIRA	cmuWiki10	A	0.112	0.212	0.400	0.157
UAmsterdam	UAMSA10mSF30	B	0.110	0.145	0.237	0.043

Table 2: Top adhoc task results ordered by ERR@20. Only the best run from each group is included in the ranking. The baseline run is discussed in the body of the paper.

Group	Run	Cat	ERR-IA@20	α-nDCG@20	NRBP	MAP-IA
msrsv	msrsv3div	A	0.347	0.491	0.303	0.068
baseline	uwgym	A	0.346	0.487	0.299	0.050
THUIR	THUIR10DvNov	A	0.336	0.474	0.289	0.070
ICTNET	ICTNETDV10R2	A	0.322	0.464	0.279	0.038
uogTr	uogTrB67xS	B	0.298	0.418	0.262	0.074
unimelb	UMd10IASF	A	0.255	0.377	0.208	0.050
CMU_LIRA	cmuWi10D	A	0.248	0.345	0.215	0.093
UAmsterdam	UAMSD10aSRfu	B	0.242	0.341	0.210	0.026
UCDSIFT	UCDSIFTDiv	B	0.210	0.312	0.170	0.062
qirdcsuog	qirdcsuog3	B	0.205	0.304	0.165	0.051

Table 3: Top diversity task results ordered by ERR-IA@20. Only the best run from each group is included in the ranking. The baseline run is discussed in the body of the paper.

6 Results

6.1 Adhoc and Diversity Tasks

Table 2 presents the top adhoc task results ordered by ERR@20. Table 3 presents the top diversity task results ordered by ERR@20. The figures mix results for both Category A and B runs. Experience with the ClueWeb09 collection suggests that the Category B subset generally contains higher quality documents than the rest of the collection. The results support this view, with several Category B runs achieving good performance.

The baseline run was created at the University of Waterloo. While the run represents an official submission to the track, it should be considered as a special case in any discussion and analysis of the track results. To create this run, the queries were submitted to a commercial search engine, and the results were filtered against the ClueWeb09 collection. Thus, the performance of this run might be considered as a very crude *lower bound* on the performance of the commercial search engine.

All runs submitted to the adhoc and diversity tasks were judged using the judging criteria of

both tasks, even runs that were not submitted to both tasks. This additional judging allows us to make direct conparisons between runs optimized for the two tasks, supporting efforts to determine if the different judging criteria and evaluation measures identify genuine differences. For example, figure 2 provides a scatter plot comparing the performance of the runs under ERR@20 and ERR-IA@20, the primary effectiveness measures for the adoc and diversity tasks respectively. While the values are correlated, there are clear differences in the relative performance of runs under the two measures.

6.2 Spam Task

Figure 3 presents receiver operating characteristic curves for the spam task submissions, along with the official baseline ("Baseline") and an additional baseline ("Britney"). The official baseline is the preliminary spam ranking supplied to the track participants. No submitted run substantially outperformed this baseline at any point on the curve, as well as by overall AUC. The additional baseline represents an essentially unsupervised spam ranking of the corpus, generated without labeled training data. Apart from the documents themselves, the only data used to generate this run was a collection of queries frequently submitted to commercial search engines (e.g., "britney spears") [7].

7 Conclusions and Future Plans

The adhoc and diversity tasks will continue in more-or-less their current format for at least one more year, with perhaps some changes in the type of topics selected. Given the small number of participants, the spam task will not be continued for TREC 2011. However, the spam labelings created by the current task will be available for the use of future participants.

For TREC 2009 and 2010, diversity/adhoc topics were chosen to be of medium-to-high frequency and ambiguous. A new direction that we are considering for TREC 2011 is to work with more obscure topics, which may still be underspecified (i.e., faceted) but may be less ambiguous. Search engines have difficulty with queries of this type, since they can rely less on click/anchor information, and popularity signals like PageRank. With these new *tough topics* we have a chance to work in an area of Web retrieval that has received relatively little attention. Given the smaller number of pages that may be relevant for these topics, we may potentially be able to create a more reusable collection, with sufficiently exhaustive judgments for the topics.

Acknowledgements

The track could not operate without the ClueWeb09 collection, created by Jamie Callan, Mark Hoy, and the Language Technologies Institute at Carnegie Mellon University. University of Waterloo graduate student Hani Khoshdel Nikkhoo generated the the track baseline runs (uwgym). We thank the participants for their hard work in making the track a success.

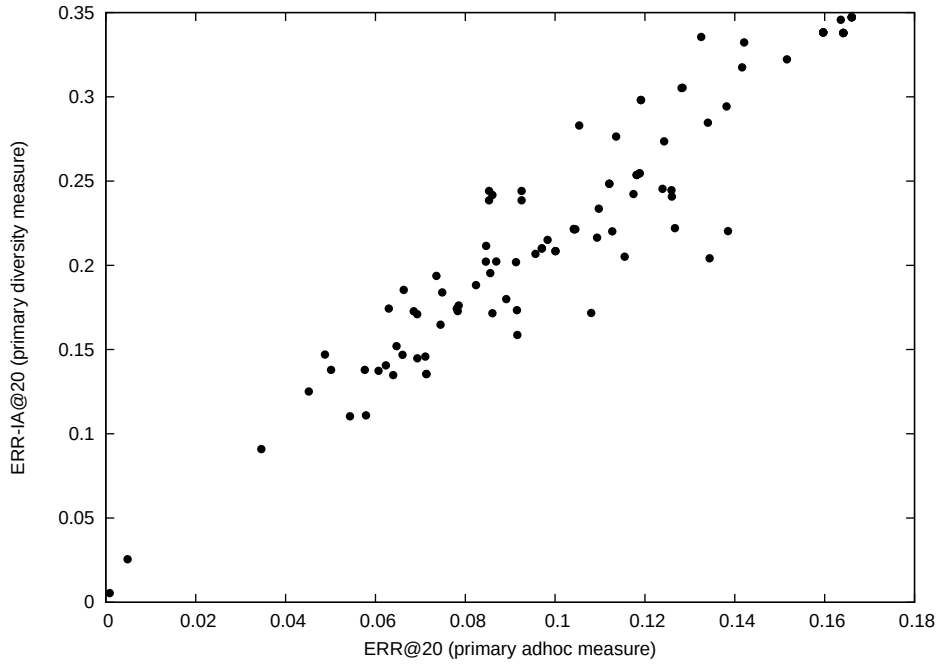

Figure 2: Comparison of runs under the primary adhoc and diversity effectiveness measures.

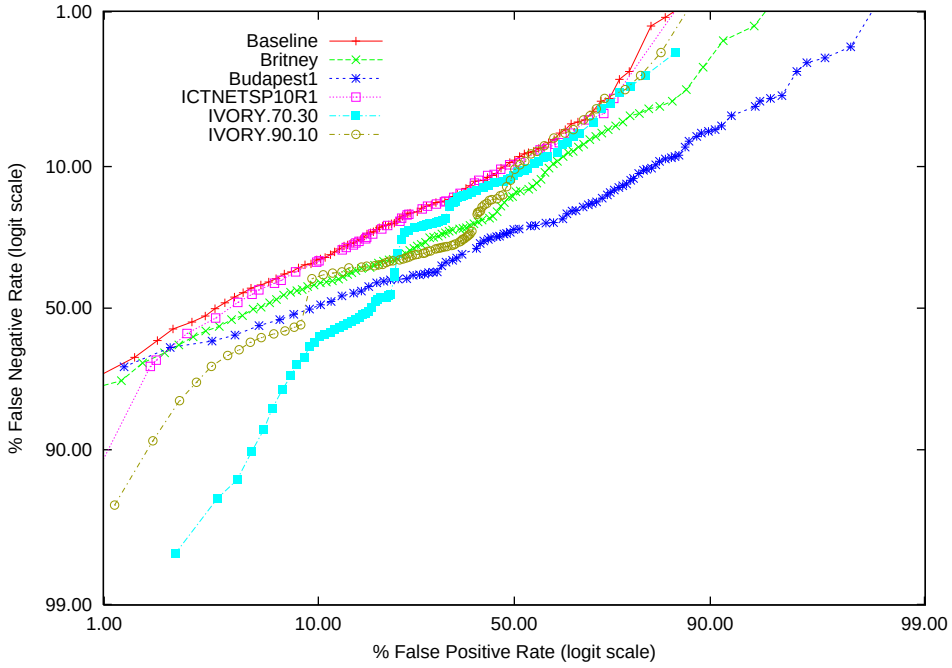

Figure 3: Web spam results. One run, Budapest2, is not plotted, since its curve is essentially identical to that of Budapest1.

References

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *2nd ACM International Conference on Web Search and Data Mining*, pages 5–14, Barcelona, Spain, 2009.

[2] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. Expected reciprocal rank for graded relevance. In *18th ACM Conference on Information and Knowledge Management*, pages 621–630, 2009.

[3] Charles Clarke, Nick Craswell, Ian Soboroff, and Azin Ashkan. A comparative analysis of cascade measures for novelty and diversity. In *4th ACM International Conference on Web Search and Data Mining*, Hong Kong, 2011.

[4] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. Overview of the TREC 2009 web track. In *18th Text REtrieval Conference*, Gaithersburg, Maryland, 2009.

[5] Charles L. A. Clarke, Maheedhar Kolla, and Olga Vechtomova. An effectiveness measure for ambiguous and underspecified queries. In *2nd International Conference on the Theory of Information Retrieval*, pages 188–199, 2009.

[6] Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkann, Stefan Büttcher, and Ian MacKinnon. Novelty and diversity in information retrieval evaluation. In *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 659–666, Singapore, 2008.

[7] Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information Retrieval*, 2011. To appear. Preprint available at `arxiv.org/abs/1004.5168`.

[8] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[9] Filip Radlinski, Martin Szummer, and Nick Craswell. Inferring query intent from reformulations and clicks. In *19th International World Wide Web Conference*, Raleigh, North Carolina, April 2010.