Overview of the TREC 2011 Web Track

Charles L. A. Clarke, University of Waterloo
Nick Craswell, Microsoft
Ian Soboroff, NIST
Ellen M. Voorhees, NIST

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 since TREC 2009, where it included both a traditional adhoc retrieval task and a new diversity task [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. For TREC 2010 the track introduced a new Web spam task and Web-style, six-level relevance assessment for the adhoc task [5]. For TREC 2011, as recommended by participants at the track planning session held during TREC 2010, we dropped the spam task but continued the other tasks essentially unchanged. As we did for TREC 2009 and TREC 2010, we based our TREC 2011 experiments on the billion-page ClueWeb09[1] collection created by the Language Technologies Institute at Carnegie Mellon University.

The tasks use a common topic set, differing only in their evaluation methodology. Topics are created from the logs of a commercial search engine, with the aid of tools developed at Microsoft Research [9]. Given a target query, these tools extract and analyze groups of related queries, using co-clicks and other information, to identify clusters of queries that highlight different aspects and interpretations of the target query. These clusters are employed by NIST for topic development. Each resulting topic is structured as a representative set of subtopics, each related to a different user need. The selection of subtopics attempts to reflect a mix of genuine user requirements for the topic.

For the adhoc task documents are judged with respect to the topic as a whole. Relevance levels are similar in structure to the levels used in commercial Web search, including a spam/junk level. Moreover, the top two levels of the assessment structure are closely related to the homepage finding and topic distillation tasks appearing in older Web Tracks. For the diversity task, documents are judged with respect to the subtopics, as well as with respect to the topic as a whole.

For TREC 2011, the topic selection process was modified slightly from previous years. For TREC 2009 and 2010, topics were chosen to be of medium-to-high frequency. TREC 2011 attempts 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 hope 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 tough topics, we may potentially be able to create a more reusable collection, with sufficiently complete judgments.

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

Task	Adhoc	Diversity	Total
Groups	14	9	16
Runs	37	25	62

Table 1: Participation in the TREC 2011 Web track

Table 1 summarizes participation in the TREC 2011 Web Track. A total of 16 groups participated in the track this year, a substantial decrease from last year, when 23 groups participated, and from TREC 2009, when 26 groups participated.

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. Figure 1 provides two examples.

Each topic contains a query field, a description field, and several subtopic fields. The query 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. 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 runs.

Each topic is assigned one of two types. Topics with ambiguous queries, such as topic 140 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 114 in the figure, have one primary interpretation, reflected in the description field. For these queries, the subtopics address various aspects of the broader topic. In all topics, the description field and the first subtopic field are identical.

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 may often have only 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 may often have a large number of relevant pages. 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="114" type="faceted">
  <query>adobe indian houses</query>
  <description>
    How does one build an adobe house?
  </description>
  <subtopic number="1" type="inf">
    How does one build an adobe house?
  </subtopic>
  <subtopic number="2" type="inf">
    information about Indian tribes that used adobe houses
  </subtopic>
  <subtopic number="3" type="nav">
    I'd like to order books or videos/CDs about how to construct adobe buildings.
  </subtopic>
</topic>

<topic number="140" type="ambiguous">
  <query>east ridge high school</query>
  <description>
    demographics of East Ridge High School in Lick Creek, Kentucky
  </description>
  <subtopic number="1" type="inf">
    demographics of East Ridge High School in Lick Creek, Kentucky
  </subtopic>
  <subtopic number="2" type="nav">
    home page for East Ridge High School in Chattanooga, Tennessee
  </subtopic>
  <subtopic number="3" type="inf">
    information about the sports program at East Ridge High School in Clermont, Florida
  </subtopic>
  <subtopic number="4" type="inf">
    description of the sports facilities at East Ridge High School in Woodbury, MN
  </subtopic>
</topic>
```

Figure 1: Examples of TREC 2011 Web track topics.

4 Methodology 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.

Group	Run	Cat	ERR@20	nDCG@20	P@20	MAP
Chinese Acad. of Sciences (ICT)	ICTNET11ADR3	A	0.157	0.283	0.345	0.175
Univ. of Glasgow (Terrier)	uogTrA45Vm	A	0.149	0.305	0.347	0.203
Univ. of Waterloo (MDS)	UWatMDSqlt	A	0.144	0.242	0.307	0.135
Microsoft Research	msrsv2011a3	A	0.143	0.273	0.327	0.172
N-A	srchvrs11b	B	0.131	0.233	0.298	0.110
Univ. of Ottawa	DFalah11	B	0.122	0.204	0.275	0.079
Univ. of Amsterdam (Kamps)	UAmsM705tiLS	B	0.119	0.202	0.273	0.085
Univ. of Waterloo (Clarke)	uwBAadhoc	A	0.119	0.172	0.230	0.079

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

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 [7], NRBP [6], 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 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 25 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.

5 Results

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.

All runs submitted to the adhoc and diversity tasks were judged according to the judging criteria of both tasks, even runs that were not submitted to both tasks. This additional judging allows us to make direct comparisons between runs optimized for the two tasks, supporting efforts to determine if the different judging criteria and evaluation measures identify genuine differences. For

Group	Run	Cat	ERR-IA@20	α-nDCG@20	NRBP
Univ. of Glasgow (Terrier)	uogTrA45Nmx2	A	0.528	0.630	0.487
Microsoft Research	msrsv2011d1	A	0.499	0.608	0.456
Univ. of Waterloo (MDS)	UWatMDSqltsr	A	0.494	0.595	0.457
Chinese Acad. of Sciences (ICT)	ICTNET11DVR3	A	0.476	0.582	0.432
Univ. of Amsterdam (Kamps)	UAmsM705tFLS	B	0.438	0.522	0.407
Univ. of Waterloo (Clarke)	uwBA	A	0.399	0.492	0.359
Univ. of Delaware (Fang)	UDCombine2	B	0.375	0.458	0.340
Centrum Wiskunde + Informatica	CWIcIA2t5b1	A	0.349	0.432	0.304

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

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 adhoc and diversity tasks respectively. While the values are correlated, there are clear differences in the relative performance of runs under the two measures.

6 Conclusions and Future Plans

Given the drop in participants, we believe that we must develop new tasks and ideas for the track to go forward for another year. We should not simply repeat the current tasks on the same collection. Some possible ideas were proposed during a workshop on Diversity in Document Retrieval held at ECIR 2011. Breakout groups at the workshop identified a number of possible extensions to the current task set, which we hope to explore during our planning session at TREC. These extensions include: 1) a query suggestion task, 2) a subtopic mining task, and 3) a snippet generation task.

Jamie Callan's research group at CMU, who created the ClueWeb09 collection, is considering the creation of a new collection. This collection will be of a similar size to ClueWeb09, but will address some of its known problems. If this new collection is available in time for TREC 2012 tasks, we plan to switch to it, providing researchers with a set of topics and judgments over the new collection.

Acknowledgements

Again this year, we extend our thanks to Jamie Callan, Mark Hoy, and the Language Technologies Institute at Carnegie Mellon University, who created and continue to distribute the ClueWeb09 collection. The track could not operate without this valuable resource.

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, Nick Craswell, Ian Soboroff, and Gordon V. Cormack. Overview of the TREC 2010 web track. In *19th Text REtrieval Conference*, Gaithersburg, Maryland, 2010.

[6] 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.

[7] 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.

[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.

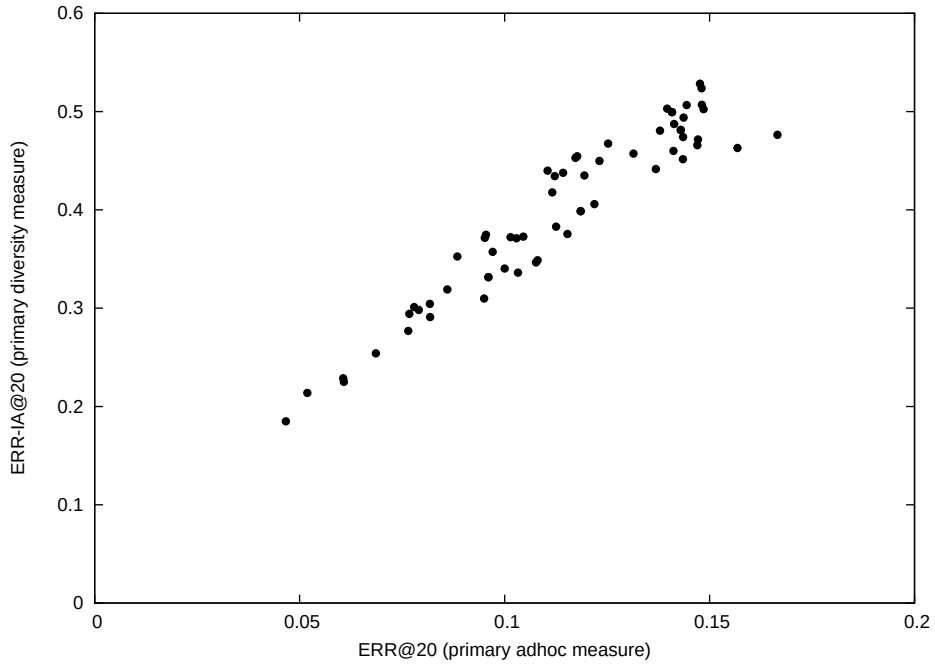

Figure 2: Comparison of runs under the primary adhoc and diversity effectiveness measures.