Overview of the TREC 2012 Web Track

Charles L. A. Clarke
University of Waterloo

Nick Craswell
Microsoft

Ellen M. Voorhees
NIST

Summary for experienced participants

If you are an experienced participant, you may not need to read the full report. Apart from the results themselves (see tables 1, 2, and 3) little has changed from TREC 2011 [6]. A six-point scale was used for relevance assessment (see section 4.1). Limitations on available assessor time meant that some topics were judged to depth 30 and others to depth 20, as well as causing other minor problems (see section 4.3). However, our plans for next year, as outlined in the concluding section, are quite different from this year.

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 [5]. For both TREC 2011 and 2012, we dropped the spam task but continued the other two tasks essentially unchanged. As we did since TREC 2009, we based our TREC 2012 experiments on the billion-page ClueWeb09[1] collection created by the Language Technologies Institute at Carnegie Mellon University.

The two 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 [10]. 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.

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

Task	Adhoc	Diversity	Total
Groups	11	8	12
Runs	28	20	48

Table 1: Participation in the TREC 2012 Web track

Table 1 summarizes participation in the TREC 2012 Web Track. A total of 12 groups participated in the track this year, a slight decrease from last year, when 16 groups participated, and a substantial decrease from 2009 and 2010, when more than 20 groups participated. One group from the University of Delaware, submitted a manual run for the diversity task; all other runs were automatic, with no human intervention at any stage.

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 162 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 155 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.

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. (*relevance grade 3*)

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

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 (i.e., 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. For binary effectiveness measures, we treat grades 1/2/3/4 as relevant and grades 0/-2 as non-relevant.

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 [9], 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:

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

```
<topic number="155" type="faceted">
  <query>last supper painting</query>
  <description>
    Find a picture of the Last Supper painting by Leonardo da Vinci.
  </description>
  <subtopic number="1" type="nav">
    Find a picture of the Last Supper painting by Leonardo da Vinci.
  </subtopic>
  <subtopic number="2" type="nav">
    Are tickets available online to view da Vinci's Last Supper in Milan, Italy?
  </subtopic>
  <subtopic number="3" type="inf">
    What is the significance of da Vinci's interpretation of the Last Supper in
    Catholicism?
  </subtopic>
</topic>

<topic number="162" type="ambiguous">
  <query>dnr</query>
  <description>
    What are "do not resuscitate" orders and how do you get one in place?
  </description>
  <subtopic number="1" type="inf">
    What are "do not resuscitate" orders and how do you get one in place?
  </subtopic>
  <subtopic number="2" type="nav">
    What is required to get a hunting license online from the Michigan Department of
    Natural Resources?
  </subtopic>
  <subtopic number="3" type="inf">
     What are the Maryland Department of Natural Resources' regulations for deer hunting?
  </subtopic>
</topic>
```

Figure 1: Examples of TREC 2012 Web track topics.

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

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

We apply `trec_eval` to compute MAP and other traditional 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, a binary judgment indicates whether or not a document satisfies the information need associated with that subtopic. For the TREC 2012 track, assessors made graded judgments. To apply evaluation measures, we mapped these graded judgments to binary judgments by treating values > 0 as relevant and values ≤ 0 as not relevant. However, the graded judgments are available in the TREC data repository for the use of interested participants.

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 [8], NRBP [7], 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. A common pool was used for both tasks, and all runs were judged 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.

We initially planned to judge all runs to depth 30. Unfortunately, judging went more slowly this year than last year, for reasons that are not clear to us. As a result, we cut back the size of the pools for 25 topics to depth 20. The topics with depth-20 pools are 152, 156, 159, 160, 161, 164, 166, 167, 169, 173, 177, 179, 181, 183, 184, 185, 188, 189, 190, 191, 192, 193, 195, 196, 198.

Even with depth-20 pools, two topics had documents remaining to be judged when available assessor time ran out. Topic 156 ("university of phoenix") had about 20 or so documents remaining, while topic 185 ("credit report") had about one-third of the documents remaining. We asked a researcher at NIST to finish judging these last two topics. We included these extra judgments in the official qrels for topics 156 and 185, but you may wish to exclude these topics when using the collection in future research, particularly if single-assessor judging is important.

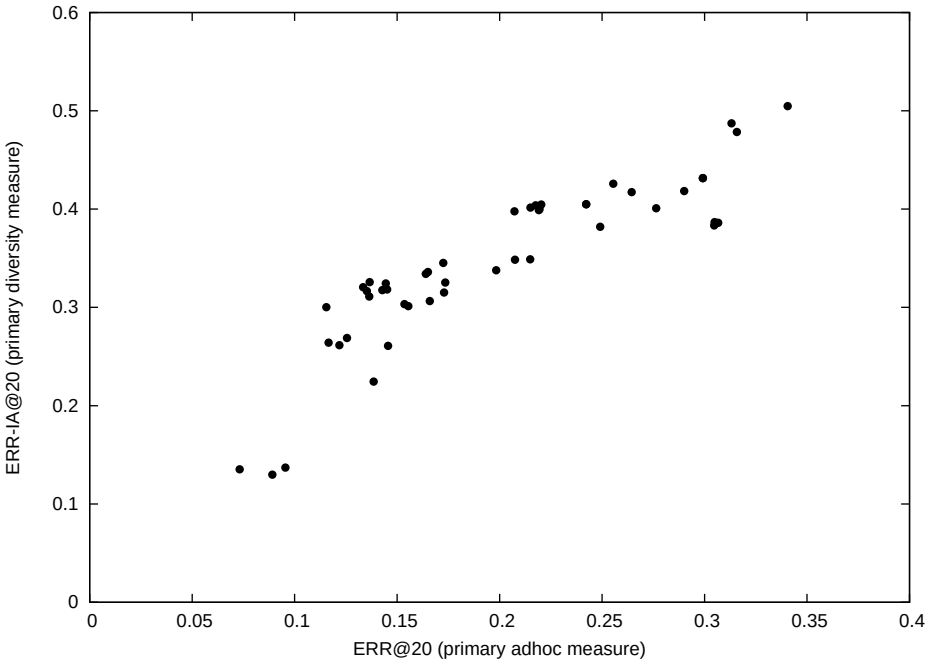

Figure 2: Comparison of runs under the primary adhoc and diversity effectiveness measures.

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 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

The Web Track will undergo a substantial change for TREC 2013. While the adhoc task will continue, we plan to drop the diversity task in favor of a new *risk-sensitive* retrieval task. This new task will explore the tradeoffs systems can achieve between effectiveness (overall gains across queries) and robustness (minimizing the possibility of significant failure, relative to a given baseline).

In addition, Jamie Callan's research group at CMU — who created the ClueWeb09 collection — have created a new ClueWeb12 collection. The size of this new collection is similar to that of ClueWeb09, but it addresses known problems with the existing collection. We plan to switch to this new collection for TREC 2013.

Group	Run	Cat	Type	ERR@20	nDCG@20	P@20	MAP
uogTr	uogTrA44xi	A	auto	0.313	0.238	0.453	0.212
srchvrs	srchvrs12c09	A	auto	0.305	0.176	0.315	0.126
uottawa	DFalah121A	B	auto	0.299	0.214	0.405	0.120
QUT_Para	QUTparaBline	B	auto	0.290	0.167	0.305	0.117
utwente	utw2012fc1	B	auto	0.219	0.113	0.221	0.061
ICTNET	ICTNET12ADR2	A	auto	0.215	0.110	0.257	0.078
IRRA	irra12c	B	auto	0.173	0.143	0.367	0.153
qutir12	qutwb	B	auto	0.166	0.146	0.308	0.131

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

Group	Run	Cat	Type	ERR-IA@20	α-nDCG@20	NRBP
uogTr	uogTrA44xu	A	auto	0.505	0.606	0.463
uottawa	DFalah121D	B	auto	0.431	0.525	0.394
utwente	utw2012c1	B	auto	0.405	0.508	0.357
srchvrs	srchvrs12c00	A	auto	0.386	0.485	0.340
ICTNET	ICTNET12DVR1	A	auto	0.326	0.422	0.280
udel	autoSTA	A	auto	0.325	0.419	0.282
LIA	lcm4res	A	auto	0.318	0.424	0.268
udel_fang	UDInfoDivSt	B	auto	0.300	0.420	0.241

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

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, Nick Craswell, Ian Soboroff, and Ellen M. Voorhees. Overview of the TREC 2011 web track. In *20th Text REtrieval Conference*, Gaithersburg, Maryland, 2011.

[7] 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.

[8] 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.

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

[10] 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.