ClueWeb09 and TREC Diversity

Charles Clarke, University of Waterloo

co-coordinators:
Nick Craswell, Microsoft
Ian Soboroff, NIST
The Big Picture

Our goal is to explore and evaluate Web retrieval technologies over a (nearly) commercial-scale collection.

2009 was our first year; continues in 2010

Tasks: adhoc, diversity, spam (new)
The collection: ClueWeb09

- One billion pages; 25TB uncompressed; multiple languages (but in practice we use only the English-language subset comprising about half of the collection)
- “Category B” subset of 50 million English pages used by many participants.
- Crawled in early 2009.

http://boston.lti.cs.cmu.edu/Data/clueweb09
Adhoc task

• Return a ranked list of documents in order of decreasing probability of relevance.
• Automatic runs only, returning the top 1000 documents for each of 50 topics.
• Judged by NIST on a four-point scale of “highly relevant”, “relevant”, “not relevant” and “not relevant, but reasonable”.

(more-or-less the same as always)
Diversity task

• Return a ranked list of documents in order of decreasing probability of relevance, where relevance is considered in the context of higher-ranked documents.
• Automatic runs only, returning the top 1000 documents for each of 50 topics.
• Judged by NIST through binary judgments with respect to sub-topics.

(closer to “real” Web retrieval?)
What does “windows” mean?

1. Microsoft Windows
   a) Can I upgrade directly from XP to Windows 7?
   b) What’s the Windows update URL?
   c) I want to download Windows Live Essentials

2. House windows
   a) Where can I buy replacement windows?
   b) What brands are available?
   c) Aluminum or vinyl?

3. Windows Restaurant, Las Vegas
50 New Topics

• Query + Detailed Subtopics
• Initial release: query only
• Based on clusters created from co-click and other information extracted from the logs of a commercial search engine (with thanks to Microsoft, Filip Radlinski, and Martin Szummer).
Raw material for topics

obama family tree
  obama(100) family(71) barack(42) tree(28) pictures(28)...
  chris(99) rock(99) video(18) quotes(9) show(9) police(9)...
  mother(100) obama(100) dunham(66) stanley(50) ann(50)...
  obama(100) s(55) father(55) barack(44) sr(33)...
  bush(100) family(85) tree(42) george(28) history(14)...

kcs
  kansas(100) kcs(100) city(100) southern(100) com(85) railroad(57)...
  kanawha(100) county(100) school(50) schools(42) calendar(21)...
  kcs(100) railroad(100) jobs(25) careers(25) career(25)...
  county(100) knox(90) schools(49) school(39) www(9) system(9)...
  energy(100) kcs(100) www(25) merger(25) kcsenergy(25) inc(25)...

Clusters used for “inspiration” (avoiding strange and rare aspects and interpretations).

Two topic types:
- faceted
- ambiguous

Two subtopic types:
- informational
- navigational
<topic number="1" type="faceted">
  <query>obama family tree</query>
  <description>
    Find information on President Barack Obama's family history, including genealogy, national origins, places and dates of birth, etc.
  </description>
  <subtopic number="1" type="nav">
    Find the TIME magazine photo essay "Barack Obama's Family Tree".
  </subtopic>
  <subtopic number="2" type="inf">
    Where did Barack Obama's parents and grandparents come from?
  </subtopic>
  <subtopic number="3" type="inf">
    Find biographical information on Barack Obama's mother.
  </subtopic>
</topic>
Topic #6: kcs

Find information on the Kansas City Southern railroad.

Find the homepage for the Kansas City Southern railroad.

I'm looking for a job with the Kansas City Southern railroad.

Find the homepage for Kanawha County Schools in West Virginia.

Find the homepage for the Knox County School system in Tennessee.

Find information on KCS Energy, Inc., and their merger with
Topics and sub-topics

- For adhoc task, relevance judged in terms of description field.
- For diversity task, relevance judged independently with respect to each subtopic.
- 3-8 subtopic per topic (mean 4.9)
## Participants

<table>
<thead>
<tr>
<th>Task Type</th>
<th>Groups submitting Category A runs</th>
<th>Groups submitting Category B runs</th>
<th>Total groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>adhoc task</td>
<td>13</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>diversity task</td>
<td>10</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>any task</td>
<td>13</td>
<td>16</td>
<td>26</td>
</tr>
</tbody>
</table>

Figure 1: Participation in the TREC 2009 Web track.
Evaluation

• **adhoc task**: Expected MAP, etc.
  – see TREC Million Query Track for details

• **diversity task**:
  – $\alpha$-nDCG, see Clarke et al. (SIGIR ’08)
  – precision-IA, see Agrawal et al. (WSDM ’09)
  – intent aware ERR, see Chapelle et al. (CIKM ‘09)
  – NRBP, see Clarke et al. (ICTIR ‘09)
  – simple measures…
results
TREC 2010

- Track continues in 2010
  - adhoc and diversity tasks continue
  - new spam task
- 50 new topics; 6-level adhoc judging
- Submissions due: August 9
  - still time to participate!
Web spam

• Simple filtering of 2009 results to remove Web spam substantially improved results.

• See http://arxiv.org/abs/1004.5168

• TREC 2010 spam task: Score each English document in the full ClueWeb09 collection according to how likely it is to be spam.
TREC 2010 Web Track

plg.uwaterloo.ca/~trecweb/2010.html
Questions?

TREC Web Track
Charles Clarke
University of Waterloo