If I Had a Million Queries

Ben Carterette, Virgil Pavlu, Evangelos Kanoulas, Javed Aslam, James Allan
TREC 2008 Million Query Track

• Traditional TREC evaluation setup
  – Depth-100 pools judged
  – 50 queries
  – Infeasible (judgment effort) and insufficient

• Million Query evaluation setup
  – Reduce judgment effort by carefully selecting
    • Documents to judge
    • Types of queries to evaluate systems on
TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?

2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?
Million Query Track Setup

8 participating sites
25 retrieval runs

10,000 Queries
GOV2

uses of alternative dispute resolution
job search vancouver washington
poem of arrival of columbus

TREC crew @ NIST
Million Query Track Setup

8 participating sites
25 retrieval runs

10,000 Queries
GOV2

uses of alternative dispute resolution
job search vancouver washington
poem of arrival of columbus

TREC crew @ NIST
Million Query Track Setup

- 10,000 Queries
- GOV2
- 8 participating sites
- 25 retrieval runs
- Retrieval results
- Assessors

TREC crew @ NIST
Million Query Track Setup

8 participating sites
25 retrieval runs

10,000 Queries
GOV2

TREC crew @ NIST

Retrieval results

Assessors

Relevance judgments
Million Query Track Setup

10,000 Queries
GOV2

TREC crew @ NIST

Retrieval results

Assessors

Relevance judgments
Document Selection and Evaluation

- Two low-cost algorithms
  - MTC (Carterette, Allan, & Sitaraman, 2006)

  Document Selection
  - Greedy on-line algorithm
  - Selects most discriminative documents
  - Targets at accurate ranking of systems

Evaluation
- Each document has a probability of relevance
- Measures as expectations over relevance distribution
Document Selection and Evaluation

• Two low-cost algorithms
  – statAP (Aslam & Pavlu, 2008)

  Document Selection
  • Stratified random sampling
  • Selects documents based on prior belief of relevance

  Evaluation
  • Apply well-established estimation techniques
  • Targets at accurate system scores
Queries

• 10,000 queries sampled from logs of a search engine.

• Queries were assigned categories
  – Long (more than 6 words) vs. Short
  – Gov-heavy (more than 3 clicks) vs. Gov-slant

<table>
<thead>
<tr>
<th></th>
<th>short</th>
<th>long</th>
</tr>
</thead>
<tbody>
<tr>
<td>gov-slant</td>
<td>2,434</td>
<td>2,434</td>
</tr>
<tr>
<td>gov-heavy</td>
<td>2,434</td>
<td>2,434</td>
</tr>
</tbody>
</table>
Judgments per Query

• Five different targets for number of judgments
  – 8, 16, 32, 64 and 128 judgments targeted
  – Equal total number of judgments per target over all queries
Relevance Judgments

• 784 of the 10,000 queries judged
• 15,211 total judgments
  – ~75% less than in past years
Relevance Judgments

- Distribution of queries per category and judgment target

<table>
<thead>
<tr>
<th>Category</th>
<th>Judgments</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-govslant</td>
<td>95</td>
<td>55</td>
<td>29</td>
<td>13</td>
<td>4</td>
<td></td>
<td>196</td>
</tr>
<tr>
<td>Short-govheavy</td>
<td>118</td>
<td>40</td>
<td>26</td>
<td>10</td>
<td>3</td>
<td></td>
<td>197</td>
</tr>
<tr>
<td>Long-govslant</td>
<td>98</td>
<td>52</td>
<td>26</td>
<td>13</td>
<td>8</td>
<td></td>
<td>197</td>
</tr>
<tr>
<td>Long-govheavy</td>
<td>92</td>
<td>57</td>
<td>21</td>
<td>14</td>
<td>10</td>
<td></td>
<td>194</td>
</tr>
<tr>
<td>Total</td>
<td>403</td>
<td>204</td>
<td>102</td>
<td>50</td>
<td>25</td>
<td></td>
<td>784</td>
</tr>
</tbody>
</table>
Evaluation Measure

• Weighted MAP:

\[ \text{wMAP} = \frac{1}{5} \sum_{j=1}^{5} \text{MAP}_j = \frac{1}{5} \sum_{j=1}^{5} \frac{1}{|Q_j|} \sum_{q \in j} \text{AP}_q \]

<table>
<thead>
<tr>
<th>Judgments</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>403</td>
<td>204</td>
<td>102</td>
<td>50</td>
<td>25</td>
<td>784</td>
</tr>
</tbody>
</table>
TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?

2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?
System Scores and Rankings

Kendall $\tau = 0.93$
TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?

2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?
Timing Info for Cost Analysis

- Query overhead

<table>
<thead>
<tr>
<th></th>
<th>refresh</th>
<th>view</th>
<th>last view</th>
<th>topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>short</td>
<td>2.34</td>
<td>18.0</td>
<td>25.5</td>
<td>67.6</td>
</tr>
<tr>
<td>long</td>
<td>2.54</td>
<td>24.5</td>
<td>31.0</td>
<td>86.5</td>
</tr>
<tr>
<td>gov-slant</td>
<td>2.22</td>
<td>22.5</td>
<td>29.0</td>
<td>76.0</td>
</tr>
<tr>
<td>gov-heavy</td>
<td>2.65</td>
<td>20.0</td>
<td>27.5</td>
<td>78.0</td>
</tr>
<tr>
<td>average</td>
<td>2.41</td>
<td>22.0</td>
<td>29.0</td>
<td>76.0</td>
</tr>
</tbody>
</table>
Timing Info for Cost Analysis

- Judging time per category and judgment target

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
<th>128</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>short</td>
<td>15.0</td>
<td>11.5</td>
<td>13.5</td>
<td>12.0</td>
<td>8.5</td>
<td>12.5</td>
</tr>
<tr>
<td>long</td>
<td>17.0</td>
<td>14.0</td>
<td>16.5</td>
<td>10.0</td>
<td>10.5</td>
<td>13.0</td>
</tr>
<tr>
<td>gov-slant</td>
<td>13.0</td>
<td>12.5</td>
<td>13.0</td>
<td>9.5</td>
<td>10.5</td>
<td>12.0</td>
</tr>
<tr>
<td>gov-heavy</td>
<td>19.0</td>
<td>13.0</td>
<td>17.0</td>
<td>12.5</td>
<td>8.5</td>
<td>13.5</td>
</tr>
<tr>
<td>average</td>
<td>15.0</td>
<td>13.0</td>
<td>15.0</td>
<td>11.0</td>
<td>9.0</td>
<td>13.0</td>
</tr>
</tbody>
</table>
Analysis of Variance

- $\sigma_s$ = variance due to systems
- $\sigma_q$ = variance due to queries
- $\sigma_{sq}$ = variance due to query-system interaction
TREC 2008 Million Query Track

- Measure the stability of

  - Scores: $\frac{\text{Variance due to systems}}{\text{Total variance}}$

  - Rankings: $\frac{\text{Variance due to systems}}{\text{Var. due to systems + Var. due to query-system}}$
MAP Variance Components

- What is the minimum cost needed to reach reliable result?
MAP Variance Components per Query Category

• Are some query types more informative than others?
Query Selection

- Are some queries types more informative than others?
TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?

2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?
Kendall’s tau Analysis

- What is the minimum cost needed to reach reliable result?
Kendall’s tau Analysis

• Are some query types more informative than others?
Kendall’s tau Analysis

• Are some query types more informative than others?
Relevance

- Percentage of relevant documents per query category and judgment target

<table>
<thead>
<tr>
<th>Category</th>
<th>Judgments 8</th>
<th>Judgments 16</th>
<th>Judgments 32</th>
<th>Judgments 64</th>
<th>Judgments 128</th>
<th>avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-govslant</td>
<td>18.7</td>
<td>12.1</td>
<td>20.2</td>
<td>13.9</td>
<td>3.0</td>
<td>14.6</td>
</tr>
<tr>
<td>Long-govslant</td>
<td>20.2</td>
<td>17.0</td>
<td>17.3</td>
<td>12.0</td>
<td>13.7</td>
<td>15.9</td>
</tr>
<tr>
<td>Short-govheavy</td>
<td>24.6</td>
<td>30.8</td>
<td>30.4</td>
<td>23.4</td>
<td>37.4</td>
<td>28.3</td>
</tr>
<tr>
<td>Long-govheavy</td>
<td>28.8</td>
<td>20.4</td>
<td>22.3</td>
<td>13.6</td>
<td>16.0</td>
<td>19.6</td>
</tr>
<tr>
<td>avg</td>
<td>23.1</td>
<td>19.3</td>
<td>22.5</td>
<td>15.2</td>
<td>15.7</td>
<td>19.3</td>
</tr>
</tbody>
</table>
TREC 2008 Million Query Track

Questions:

1. Can low-cost methods reliably evaluate retrieval systems?

2. What is the minimum cost needed to reach reliable result?

3. Are some query types more informative than others?

4. Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?
Cost-Benefit Analysis

• Is it better to judge a lot of documents for a few queries or a few documents for a lot of queries?

![Graph showing the relationship between assessor time (in hours) and the number of judgments per query.](Image of graph)

- 64 judgments & 50 queries
Conclusion

• Low-cost methods reliably evaluate retrieval systems with very few judgments

• Minimum cost to reach reliable results
  – 10-15 hours of judgment time

• Some queries more informative than others
  – Gov-heavy more informative than gov-slant

• 64 judgments per query with around 50 queries is optimal for assessing systems’ performance ranking