Low Cost Evaluation for Diversity and Novelty

Kai Hui
Max-Planck-Institut für Informatik, D5
khui@mpi-inf.mpg.de
Outline

- Background
- Motivation
- Preliminary Results
- Related Work
- Our Method
- Next Steps
Background:
Diversity and Novelty Evaluation
TREC Diversity and Novelty Example

Query: No. 196 from WT2012

Query: sore throat
Subtopic 1: What causes a sore throat?
Subtopic 2: Find home remedies for a sore throat.
Subtopic 3: Find information on throat cancer.
Subtopic 4: What does it mean when my throat is sore on only one side?

<table>
<thead>
<tr>
<th>Query-id</th>
<th>Subtopic</th>
<th>Docid</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>196</td>
<td>1</td>
<td>clueweb09-enwp02-06-01125</td>
<td>1</td>
</tr>
<tr>
<td>196</td>
<td>2</td>
<td>clueweb09-enwp02-06-01125</td>
<td>0</td>
</tr>
<tr>
<td>196</td>
<td>3</td>
<td>clueweb09-enwp02-06-01125</td>
<td>1</td>
</tr>
<tr>
<td>196</td>
<td>4</td>
<td>clueweb09-enwp02-06-01125</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>......</td>
</tr>
</tbody>
</table>
Diversity and Novelty Measure Example: ERR-IA

\[ ERR - IA = \sum_{top-k} \frac{1}{k} \sum_{subtopic \ i} g_{k,i} (1 - \alpha)^{c(k,i)} \]

- \( g_{k,i} \): relevance labels for the \( k \)-th document on subtopic \( i \)
- \( c(k, i) \): the count of relevant documents for subtopic \( i \) before the \( k \)-th document
- \( \alpha \): the parameter for penalizing the repeating subtopics, normally set as 0.5

- Apart from ERR-IA, \( \alpha \)-nDCG and NRBP are also popular measures.
- Their definition is directly based on the relevance labels, thus the evaluation quality highly depends on the labels.
Datasets

Document Collections
- ClueWeb09 Category A (CwA): 500 M English web pages
- ClueWeb09 Category B (CwB): 50 M English web pages
- Constructed Dataset (CwC): 450 M web pages from CwA but not in CwB

Query Sets and Labels
- TREC Web Track (WT) 2009-2012, 200 queries with their labeled documents

Objective
- Reconstruct the ranking of runs according to ERR-IA with incomplete judgments
- Kendall’s $\tau$ is used to measure the correlation among rankings
- Kendall’s $\tau$: scaled between -1 and 1

\[
\tau = \frac{(Number\ of\ Concordant\ Pairs)-(Number\ of\ Discordant\ Pairs)}{Total\ Number\ of\ Pair\ Combinations}
\]
Motivation:
Towards Lost Cost Evaluation
Cost of Evaluation

- Top-k pooling: collect top-k from all candidate runs to generate a pool
- Manually label every subtopic document pair in the pool
- Limit to shallow pooling depth, e.g., top-25

<table>
<thead>
<tr>
<th>Year</th>
<th>#Query</th>
<th>#Systems</th>
<th>Pooling size</th>
<th>#Total labeled doc</th>
<th>#Labeled Relevant doc</th>
</tr>
</thead>
<tbody>
<tr>
<td>2009</td>
<td>50</td>
<td>48</td>
<td>20</td>
<td>24,817</td>
<td>4,942</td>
</tr>
<tr>
<td>2010</td>
<td>50</td>
<td>32</td>
<td>20</td>
<td>15,352</td>
<td>6,553</td>
</tr>
<tr>
<td>2011</td>
<td>50</td>
<td>62</td>
<td>25</td>
<td>19,344</td>
<td>5,030</td>
</tr>
<tr>
<td>2012</td>
<td>50</td>
<td>48</td>
<td>20/30</td>
<td>16,036</td>
<td>5,559</td>
</tr>
</tbody>
</table>
Low Cost Evaluation Framework

Three Components in Evaluation:

Selects fewer candidate documents to label
- Better discriminating ability, e.g., only few systems can return a certain relevant document
- Based on document content, e.g., centroid documents in the similarity space are more representative
- Combine the above two methods

Reuses existing labels.
- Regard unlabeled as irrelevant
- Remove unlabeled documents
- Predict labels for the unlabeled documents

Test Query Set

Manual Labeling

Document Collections
Preliminary Results:
Label Fewer Documents & Reuse the Labels
Randomly Select Fewer Documents to Label

- Percentage of available labels with random sampling versus the Kendall’s tau correlation, repeating 30 times: **evaluation with 60% of labels is not reliable**
- Lost cost evaluation can’t be realized by simply random sampling, and the evaluation is sensitive to the completeness of the labels
How to Select?

Observation: Relevant Documents Distribute not Uniformly

- Left: percentage of relevant (red) / total (blue) w.r.t. the pooling depth
- Right: percentage of total labels w.r.t. the relevant labels
- Relevant documents distribute not uniformly on the pooling depth, i.e., shallow pool contain larger portion of the relevant documents
Select According to the Rank from Rival Systems

Preliminary Results: Labels on Shallow Pool Simulate Complete Evaluation Perfectly

- Kendall’s tau correlation between the ranking determined by full judgment w.r.t. the judgment on different pooling depth, for measures ERR-IA, $\alpha$-nDCG and the NRBP
- Label on top-6 pool (40% of total label) is enough to get over 0.9 correlation w.r.t. complete judgment
- We can also use document content or discriminating ability etc.
Difficulties in Reusing the Labels: Unlabeled Documents

How many do we miss? (Zobel, 98)

- Count the number of new relevant documents in pool@i given labeled pool@i-1
- Fit the curve for existing pool and predict for deeper pool, e.g., pool@100
- Pool@20 covers 25% - 30% relevant documents
- Unlabeled relevant documents are due to their low rank, e.g., rank at 50

Why do the unlabeled documents exist?

- Incomplete coverage of the relevant documents in existing judged pool.
Existence of the Unlabeled Documents: Predict the Labels

When might they exist?
- When evaluates systems not included in the pooling
- On a new document collection
- When going deeper than the pooling depth, e.g., ERR-IA@30

Why are they problematic?
- Direct dependence on the labels, e.g., ERR-IA
- Missing labels have to be mitigated before evaluation

How to deal with them?
- Regard unlabeled as irrelevant: underestimates (Sakai et.al, 12)
- Remove unlabeled documents (condensed list): overestimate (Sakai et.al, 12)
- Predict the labels for unseen documents (Büttcher et.al, 07)(Carterette & Allan, 07)
Reuse the Existing Labels

Leave N Out: Evaluate Systems without Contribution to the Pool

- Percentage of certain percentage of systems versus the Kendall’s tau correlation, repeating 30 times: evaluation with less than 50%-60% systems contributing labels is not reliable
- The existing measurement is not reusable, being biased towards systems without contributions to the pool
Brief Review of Related Works
Related Work for Low Cost Evaluation

- **Sample documents to be labeled:**
  Identify the crucial documents to label by selecting documents with best discriminating ability, like MTC. (Carterette et. al, 06)

- **Learning to predict the missing labels**
  Mitigate the missing labels by predicting labels. Only has been tested for mitigating small portion of missing labels on adhoc task.
  (Büttcher, 07) (Carterette & Allan, 07)

- **Condensed list of relevance judgment**
  Remove all the unjudged documents instead of regarding them as irrelevant, tend to over-estimate the unlabeled systems. (Sakai, 13)

- **Reduces query numbers:**
  Use fewer queries for testing and conclude statistical significant result. Most are retrospective method, the performance is unclear. (Robertson, 11)
Our Works on Reusable Evaluation:
Results for Pointwise Prediction & Listwise Prediction
Mitigate the Unlabeled Documents: Predict the Missing Labels

Given query, subtopics, and top-k documents to evaluate.

Options for prediction:

- **Pointwise prediction** (Büttcher et.al 07) (Carterette & Allan, 07)
  
  <query, subtopic, doc>: <relevant, irrelevant>

- **Pairwise prediction** (A promising method, future work):
  1) <query, doc_1>: <relevant, irrelevant>
     <query, doc_2>: <relevant, irrelevant>
  1) <query, subtopics, doc_1, doc_2>: <same subtopic, different subtopics>

- **Listwise prediction**
  <query, subtopics, top-k docs>: the diversity and relevance of top-k docs
Results for Pointwise Prediction
Predict Labels for Document Subtopic Pair

Prediction Setting
- Given information of query, subtopic and the tf-idf for each documents
- Select terms with the decreasing order of the collection frequency
- Similar setting with (Büttcher et.al 07)

Learning Tools
- Scikit package based on python (http://scikit-learn.org/)
- Use Naïve Bayern, Linear Regression and the SVM to train
- With default setting of the parameters from the toolkit

Performance
- Partially mitigate the missing labels, can reconstruct the ranking of systems with as less as 40%-50% available labels
- ? Why we need listwise prediction…….
Percentage of available labels with random sampling versus the Kendall’s tau correlation, repeating 6 times: **evaluation with 40%-50% of labels is not reliable**

- Pointwise prediction can partially mitigate the missing labels
Listwise Prediction
Reusable Measure Framework & Current Method

Generate subtopic distributions

Given query and its t subtopics:
Merge all corresponding relevant documents to generate the LM for each subtopic.

Compute distance matrix against the subtopic distributions

Given top-k documents to evaluate
Compute KL-divergence at each position i, e.g., top-i, and for each subtopic. Get a k×t distance matrix.

Compute measure based on the distance matrix

Given distance matrix:
Summarize the distance matrix $D_{k,t}$ with simple function, i.e., Abs and Delta.
**Listwise Prediction**

**Describe the Shift & Map the Shift to Measure**

- **KL-divergence.** On each subtopic $i$, compute the divergence between the subtopic LM $\Theta_q$ and the top-$k$ LM $\Theta_R$ at each position $k$, computing the divergence matrix to describe the shift procedure.

$$KLD(\Theta_q \| \Theta_R) = \sum_{v \in V} P[v|\Theta_q] \log \left( \frac{P[v|\Theta_q]}{P[v|\Theta_R]} \right)$$

- Other options to replace $distance(top - k, t - th subtopic)$: use the probability to compute the $D_{k,t}$

- Final step: map each $D_{k,t}$ to a effective measure, i.e., a real value
Evaluation: Robustness with Incomplete Labels

Percentage of Available Labels VS Correlation for 4 years

X-axis: Percentage of labels  
Y-axis: Correlation against system ranking with complete labels
Evaluation: Reuse Labels to Evaluate Unlabeled System

- **Leave One Run Out (LORO):** given N runs, gather labels from N-1 runs, regard the remaining run as missing run
- **Measure Difference:** compute the difference of ERR-IA score of the missing run on complete and LORO judgments
- **Overestimate & Underestimate:** the near zero difference indicates the measure has no bias in evaluating unlabeled run
## Evaluation: Reuse Labels on New Document Collection

### Kendall’s $\tau$ Correlation with Label Based Measures

<table>
<thead>
<tr>
<th></th>
<th>AbsNb</th>
<th>AbsRb</th>
<th>DeltaNb</th>
<th>DeltaRb</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2009</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$-nDCG</td>
<td>.72</td>
<td>.70</td>
<td>.73</td>
<td>.69</td>
</tr>
<tr>
<td>ERR-IA</td>
<td>.78</td>
<td>.78</td>
<td>.76</td>
<td>.76</td>
</tr>
<tr>
<td>NRBP</td>
<td>.74</td>
<td>.74</td>
<td>.70</td>
<td>.74</td>
</tr>
<tr>
<td><strong>2010</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$-nDCG</td>
<td>.72</td>
<td>.66</td>
<td>.74</td>
<td>.76</td>
</tr>
<tr>
<td>ERR-IA</td>
<td>.70</td>
<td>.66</td>
<td>.72</td>
<td>.75</td>
</tr>
<tr>
<td>NRBP</td>
<td>.68</td>
<td>.65</td>
<td>.71</td>
<td>.73</td>
</tr>
<tr>
<td><strong>2011</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$-nDCG</td>
<td>.71</td>
<td>.76</td>
<td>.79</td>
<td>.81</td>
</tr>
<tr>
<td>ERR-IA</td>
<td>.67</td>
<td>.75</td>
<td>.73</td>
<td>.81</td>
</tr>
<tr>
<td>NRBP</td>
<td>.64</td>
<td>.74</td>
<td>.69</td>
<td>.79</td>
</tr>
<tr>
<td><strong>2012</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$-nDCG</td>
<td>.23</td>
<td>.40</td>
<td>.31</td>
<td>.51</td>
</tr>
<tr>
<td>ERR-IA</td>
<td>.26</td>
<td>.44</td>
<td>.31</td>
<td>.54</td>
</tr>
<tr>
<td>NRBP</td>
<td>.26</td>
<td>.45</td>
<td>.31</td>
<td>.54</td>
</tr>
</tbody>
</table>

Generate subtopic distribution on CWB and evaluate runs on CWC
Next Steps:

- Reusable Evaluation
- Select Documents to Label
Future Works in Low Cost Evaluation

Reusable Evaluation:

- Generation of subtopic distribution: text summarization, snippet generation etc..
- Regression: from distance matrix to the value of effectiveness measure
  \[ \text{Measure Score} = f(Abs, Delta) \]
- Pairwise/Pointwise prediction to predict the missing labels

Select Documents to Label:

- Select documents combining the discriminating ability of the documents and of the document content
Reference

- Justin Zobel: How reliable are the results of large-scale information retrieval experiments?, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.307-314, August 24-28, 1998, Melbourne, Australia.


- Ben Carterette, James Allan, Ramesh Sitaraman: Minimal test collections for retrieval evaluation, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, August 06-11, 2006, Seattle, Washington, USA

- Robertson, S: On the contributions of topics to system evaluation. In European conference on information retrieval (pp. 129–140).

Q & A