Dealing with Incomplete Judgments in Cascade Measures

Kai Hui¹, Klaus Berberich¹, Ida Mele²
¹Max Planck Institute for Informatics
{khui, kberberi}@mpi-inf.mpg.de
²Università della Svizzera italiana (USI)
Ida.mele@uci.ch

The 3rd ACM International Conference on the Theory of Information Retrieval
Manual judgments in the evaluation for diversification are especially expensive, due to the judgments are based on individual subtopics.

Selectively label fewer documents or reuse existing judgments is desirable, but both lead to incomplete judgments.

Established measures require complete judgments for an accurate evaluation, and only consume manual labels (e.g., -1, 0, 1, 2…)
Background

IR Evaluation Pipeline Revisit

Test Queries

System 1
System 2
System 3

Document Relevance: Binary or Graded

No Document Similarity, e.g., MAP

System 1
System 2
System 3

With Document Similarity, e.g., ERR-IA

System 1
System 2
System 3

Collect Documents

Manual Assessment

Compute Measures to Rank Systems
Background
Revisit the Evaluation for Diversification

Rain Man

Subtopic 1. Where can I watch the full “Rain Main” movie online?
Subtopic 2. Find information about the real person on which the Rain Man movie is based.
Subtopic 3. Find movie reviews of “Rain Man”.
Subtopic 4. Find quotes from the “Rain Man” movie.
Subtopic 5. Find the lyrics to Eminem’s “Rain Man”.

Manual assessors provide judgments for individual subtopics, such as document $d_1$ is judged as relevant (label 1) to subtopic 1.
Existing works successfully catered for incomplete judgments when few are missing (less than 50%)

We attempt to develop measures that can accurately evaluate on more sparse judgments, namely, when missing more than 50% judgments

Fully utilize the judgments by employing the content of documents beyond the labels
Objective

Develop novel evaluation measures for diversification, consuming sparse judgments.
Method Overview

- Represent each subtopic with a language model (LM) based on manual judgments
- Represent a ranking by a series of language models (LM) estimated for top-1, top-2, ..., top-k documents
- Compute the divergence between individual language models over each position for individual subtopics
- Convert the matrix of divergence into a scalar
Method

(1) Represent Subtopics with LMs

Query

$\text{Subtopic}_1 \rightarrow \text{LM}(S_1)$

$\text{Subtopic}_2 \rightarrow \text{LM}(S_2)$

$\vdots$

$\text{Subtopic}_k \rightarrow \text{LM}(S_k)$
Method
(2) Represent Ranking with Cascade of LMs

DOC_1 \rightarrow TOP-1 \rightarrow LM(TOP-1)

DOC_2 \rightarrow TOP-2 \rightarrow LM(TOP-2)

DOC_3 \rightarrow TOP-3 \rightarrow LM(TOP-3)

DOC_4 \rightarrow TOP-4 \rightarrow LM(TOP-4)

DOC_5 \rightarrow TOP-5 \rightarrow LM(TOP-5)
Method

(3) KL-Divergence for Evaluation

<table>
<thead>
<tr>
<th></th>
<th>LM(TOP-1)</th>
<th>LM(TOP-2)</th>
<th>LM(TOP-3)</th>
<th>LM(TOP-4)</th>
<th>...</th>
<th>LM(TOP-k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LM(S₁)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM(S₂)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LM(Sᵢ)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
g(i,k) = \max(0, 1 - \frac{\text{KLD}(\text{LM}(S_i) \mid \text{LM(TOP-k)})}{\text{KLD}(\text{LM}(S_i) \mid \text{LM(D)})})
\]

KLD between the LM of TOP-k and the LM of individual subtopics

Normalization factor: relative to a background language model, e.g., a collection language model.

While going deeper in a ranking, the gains of relevant information are computed relative to different subtopics at each position.
Method

(3) KL-Divergence for Evaluation

Gains over Different Positions Relative to Different Subtopics

- Subtopic 1
- Subtopic 2
- Subtopic 3
- Subtopic 4

Relative Gains

Top-k

top-1, top-2, top-3, top-4, top-5, top-6, top-7, top-8, top-9, top-10
### Method

(4) Summarize the KL-Divergence

- At each position, among different subtopics employ the maximum of the absolute gains *(Abs)* or of the gain difference relative to the previous position *(Delta)*
- Sum up the gains over individual positions after rank-biased normalization *(RB)*
Evaluation
Experimental Setting

- **Dataset**
  TREC Web Track 2011–2014, 64 k labeled documents, 200 queries

- **Simulate incomplete judgments**
  Randomly sample p% judgments as incomplete judgments

- **Baselines**
  Treat unlabeled documents as non-relevant or non-existent

- **Benchmark**
  Kendall’s τ correlation: approximation to the system ranking under standard measures with complete measures
Evaluation
Results

- Results on four years
- x-axis represents the percentage of the available judgments; y-axis is the correlation
- Dashed black curves are established measures (baselines)
- Red/blue curves represent two variants of the proposed measures
- Proposed measures are robust after more than 10% judgments are available
- However, it is hard for the proposed measures to get beyond 0.9 correlation
Thank You!

contact: khui@mpi-inf.mpg.de