**Aims**
- Develop a time-aware evaluation paradigm for streaming collections
  - Capture how retrieval effectiveness changes over time
  - Deal with ground truth of bursty nature
  - Accommodate various underlying user models
  - Test the ideas on CCR

**Evaluation methodology**
- Slicing time
  - Measure slice relevance
- Aggregating slice relevance

**CCR @TREC 2012 KBA**
- Cumulative citation recommendation
  - Filter a time-ordered corpus for documents that are highly relevant to a predefined set of entities
  - For each entity, provide a ranked list of documents based on their “citation-worthiness”
- Evaluation metrics are set-based (using a confidence cut-off)

**CCR @TREC 2012 KBA**
- Cumulative citation recommendation
  - Filter a time-ordered corpus for documents that are highly relevant to a predefined set of entities
  - For each entity, provide a ranked list of documents based on their “citation-worthiness”
  - Evaluation metrics are set-based (using a confidence cut-off)

**Results are evaluated in a single batch** (temporal aspects not considered)
Overview

1. Slicing time

- Simplifying assumptions
  - Slices are non-overlapping
  - Unconcerned about slices that don’t contain any relevant documents

(A) Uniform slicing
  - Slices of equal length

(B) Non-uniform slicing
  - Slices of varying length

Overview

2. Measuring slice relevance

- Ranked list of documents within a given slice
  \[ d = <d_1, \ldots, d_n> \]

- Evaluation metric
  \[ m(d_i, q) \]
  - Standard IR metrics
    - MAP, R-Prec, NDCG

Overview

3. Aggregating slice relevance

- Probabilistic formulation to estimate the likelihood of relevance
  \[
P(r = 1|d, q, m) = \sum_{i \in I} P(r = 1|d_i, q, i)P(i|q)
\]

Slice importance

- Uniform slicing
  - All slices are equally important
  \[ P(i|q) = \frac{1}{7} \]

- Non-uniform slicing
  - Bursty periods (i.e., slices with more relevant documents) are more important
  \[ P(i|q) = \frac{\#R(i, q)}{\sum_{i \in I} \#R(i, q)} \]

Experiments

- Official TREC 2012 KBA CCR runs
  - 8 systems, best run for each system
  - Only uniform time slicing
  - Binary relevance
Results
Atemporal vs. temporal ranking (MAP, weekly slicing)

Results
Atemporal vs. temporal ranking (MAP, daily slicing)

Zooming in

Findings
- Top performing teams are (almost) always the same, independent of the metric
- Temporal evaluation provides additional insights

Wrap-up
- Framework for temporal evaluation
- Applied to the evaluation of TREC 2012 KBA CCR systems
- Future work
  - Non-uniform slice weighting
  - Other streaming tasks/collections (e.g., microblog search)
  - Generalize to other time-aware information access tasks

Questions?

Online appendix:
http://ciir.cs.umass.edu/~dietz/streameval/