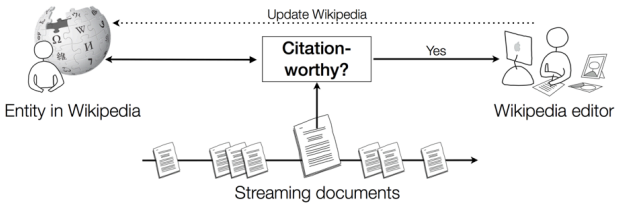
Time-aware Evaluation of Cumulative Citation Recommendation Systems

Laura Dietz, Jeffrey Dalton
CIIR, University of Massachusetts, Amherst

Krisztian Balog
University of Stavanger

SIGIR 2013 workshop on Time-aware Information Access (#TAIA2013) | Dublin, Ireland, Aug 2013

CCR @TREC 2012 KBA



Evaluation methodology

Target entity: Aharon Barak

	urlname	stream_id	score
Positive	Aharon_Barak	1328055120-f6462409e60d2748a0adef82fe68b86d	1000
	Aharon_Barak	1328057880-79cdee3c9218ec77f6580183cb16e045	500
	Aharon_Barak	1328057280-80fb850c089caa381a796c34e23d9af8	500
	Aharon_Barak	1328056560-450983d117c5a7903a3a27c959cc682a	480
	Aharon_Barak	1328056560-450983d117c5a7903a3a27c959cc682a	450
	Aharon_Barak	1328056260-684e2f8fc90de6ef949946f5061a91e0	430
	Aharon_Barak	1328056560-be417475cca57b6557a7d5db0bbc6959	428
	Aharon_Barak	1328057520-4e92eb721bfbfdfa0b1d9476b1ecb009	428
			Cutoff
Negative	Aharon_Barak	1328058660-807e4aaeca58000f6889c31c24712247	380
	Aharon_Barak	1328060040-7a8c209ad36bbb9c946348996f8c616b	380
	Aharon_Barak	1328063280-1ac4b6f3a58004d1596d6e42c4746e21	375
	Aharon_Barak	1328064660-1a0167925256b32d715c1a3a2ee0730c	315
	Aharon_Barak	1328062980-7324a71469556bcd1f3904ba090ab685	263

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"

CCR @TREC 2012 KBA

- Cumu
  - 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"

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

CCR @TREC 2012 KBA

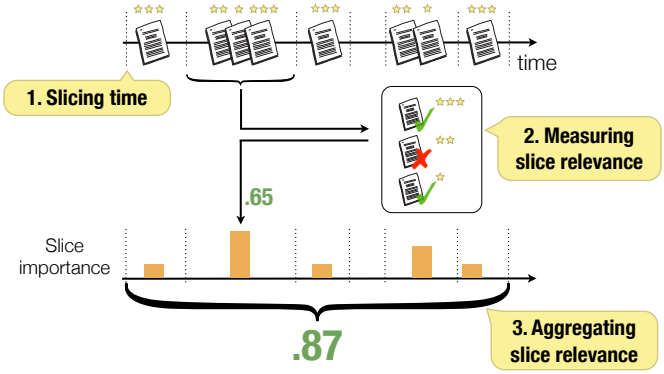
- Cumulative citation recommendation
  - Filter a **time-ordered corpus** for documents that are highly relevant to a predefined set of entities
  - For each entity, provide a **ranked list of documents** based on their "

Evaluation metrics are set-based
(using a confidence cut-off)

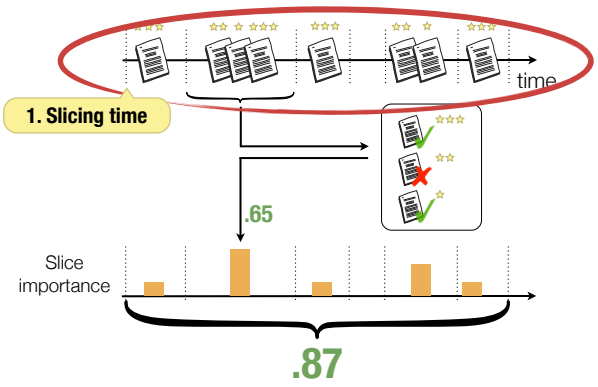
Aims

- 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

Overview



Overview

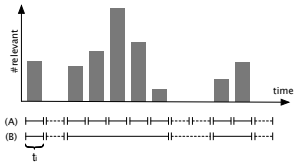


Slicing time

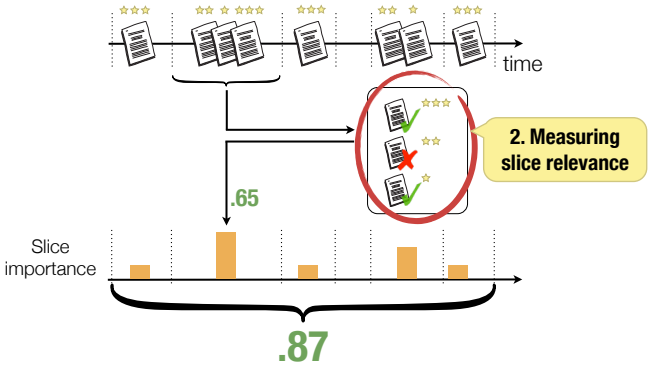
- Simplifying assumptions
  - Slices are non-overlapping
  - Unconcerned about slices that don't contain any relevant documents

(A) Uniform slicing
- Slices of equal length

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



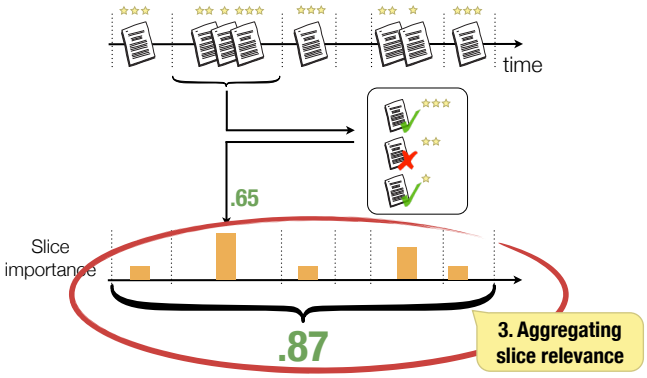
Overview



Measuring slice relevance

- Ranked list of documents within a given slice
  $\mathbf{d} =< d_1, \ldots, d_n >$
- Evaluation metric
  $m(\mathbf{d}_i, q)$
  - Standard IR metrics
    - MAP, R-Prec, NDCG

Overview



Aggregating slice relevance

- Probabilistic formulation to estimate the likelihood of relevance

$$P(r = 1|\mathbf{d}, q, m) = \sum_{i \in I} \underbrace{P(r = 1|\mathbf{d}_i, q, i)}_{\text{Slice-based relevance}} \underbrace{P(i|q)}_{\text{Slice importance}}$$

$$\approx m(\mathbf{d}_i, q)$$

Slice importance

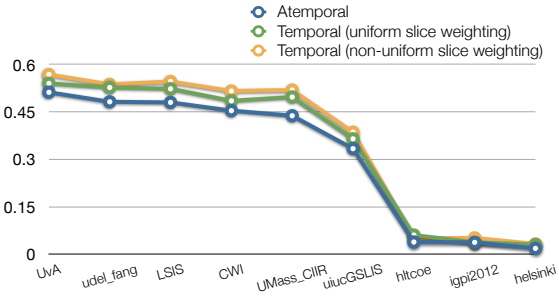
- Uniform slicing
  - All slices are equally important
    $P(i|q) = \frac{1}{I}$
- Non-uniform slicing
  - Bursty periods (i.e., slices with more relevant documents) are more important
    $P(i|q) = \frac{\#R(i,q)}{\sum_{i \in I} \#R(i,q)}$

Experiments

- Official TREC 2012 KBA CCR runs
  - 8 systems, best run for each system
- Only uniform time slicing
- Binary relevance

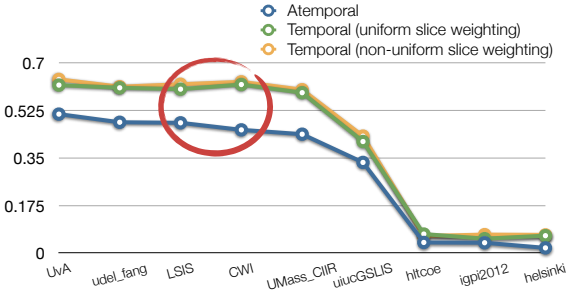
Results

Atemporal vs. temporal ranking (MAP, weekly slicing)

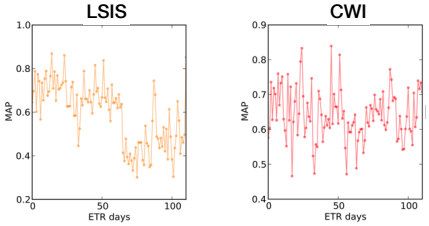


Results

Atemporal vs. temporal ranking (MAP, daily slicing)



Zooming in



	atemporal (MAP)	temporal (MAP)			
		weekly slicing		daily slicing	
		uniform	non-uniform	uniform	non-uniform
LSIS	**0.48**	**0.52**	**0.54**	0.60	0.62
CWI	0.45	0.48	0.51	**0.62**	**0.63**

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/