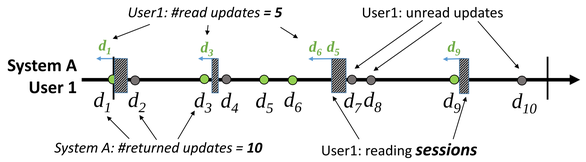
Pooling for User-Oriented Evaluation Measures

GAURAV BARUAH, ADAM ROEGIEST, MARK D. SMUCKER {gbaruah, aroegies, mark.smucker}@uwaterloo.ca

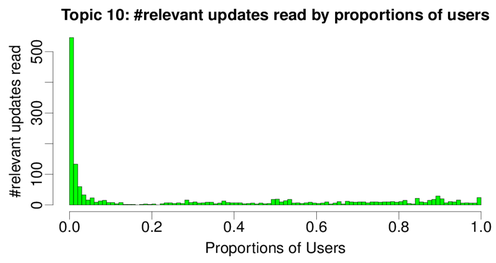
UNIVERSITY OF WATERLOO

BACKGROUND: Consider a **user model** for **streaming information access**[MSU 2015] (e.g. a stream of news updates filtered from the web).



The frequency and duration of ***sessions*** depends on the user's **level of interest** and **availability of time**. The number of updates read in a session depends on the user's **reading speed**[TWS 2014].

OBSERVATION: Document **score-based pooling** helps to identify relevant updates, however, many identified relevant updates are **not read by** (modeled) **users.**



We simulated 10,000 users reading updates from the runs submitted to the Temporal Summarization Track (TST) 2013[TST 2013].

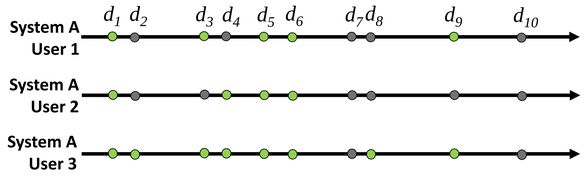
For topic 10, 1616 relevant updates were returned across all runs.

- only 5 were read by all modeled users.
- 25 were read by > 99% of users.
- 546 were read by < 1% of the users.

Other topics at TST 2013 have similar proportions of read relevant updates.

QUESTION: Can we construct an evaluation pool using updates that have a higher chance of being read?

1) $P(d_i)$: Probability of document i being read



Number of updates read across all users = 17

Balanced probabilities

$P(d_1) = \frac{1}{3}\left(\frac{1}{5} + \frac{1}{4} + \frac{1}{8}\right) = 0.192$

$P(d_2) = \frac{1}{3}\left(0 + 0 + \frac{1}{8}\right) = 0.041$

$P(d_3) = \frac{1}{3}\left(\frac{1}{5} + 0 + \frac{1}{8}\right) = 0.108$

Averaged across all users.

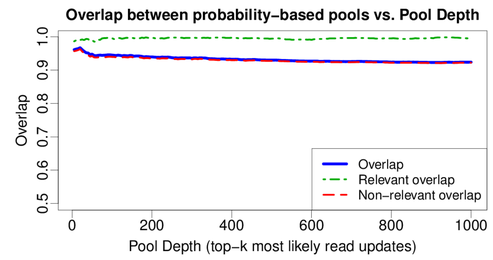
Unbalanced probabilities

$P(d_1) = \left(\frac{3}{17}\right) = 0.176$

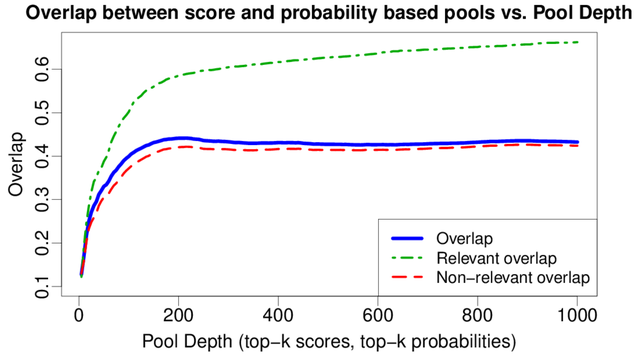
$P(d_2) = \left(\frac{1}{17}\right) = 0.058$

$P(d_3) = \left(\frac{2}{17}\right) = 0.117$

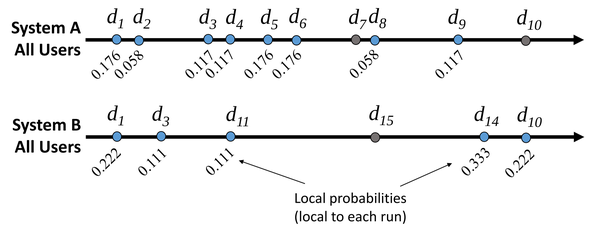
Favors users who read more.



2) Depth Pooling: top-k scores vs probabilities



3) Alternative pooling: Probability mass cover



Local probability mass cover = 0.5 (local to each run)

System A: $P(d_1) + P(d_5) + P(d_6) = 0.176 + 0.176 + 0.176 = 0.528$

System B: $P(d_{14}) + P(d_1) = 0.333 + 0.222 = 0.555$

→ **Pool = $\{d_1, d_5, d_6, d_{14}\}$**

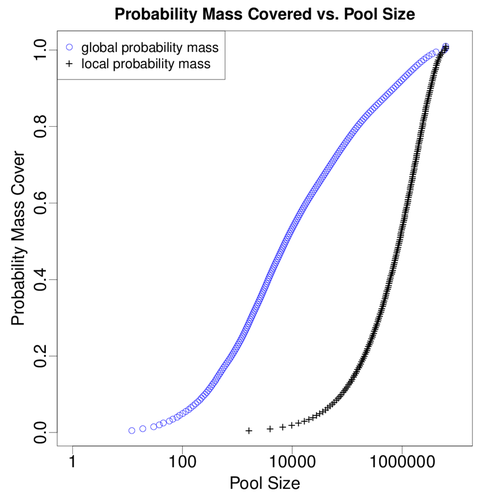
Global probability mass cover = 0.5 (across all runs)

$P(d_1) + P(d_{14}) + P(d_3) + P(d_{10})$

$= \frac{0.176+0.222}{2} + \frac{0+0.333}{2} + \frac{0.117+0.111}{2} + \frac{0+0.222}{2}$

$= 0.199 + 0.167 + 0.114 + 0.111 = 0.591$

→ **Pool = $\{d_1, d_3, d_{10}, d_{14}\}$**



KEY REFERENCES:

[TST 2013] Aslam, Diaz, Ekstrand-Abueg, Pavlu and Sakai, *TREC 2013 Temporal Summarization*, TREC 2013.

[MSU 2015] Baruah, Smucker and Clarke, *Evaluating Streams of Evolving News Events*, SIGIR 2015.

[TWS 2014] Clarke and Smucker, *Time Well Spent*, IIiX 2014.

CONCLUSIONS: Score-based and probability-based depth pooling result in different pools. Given probabilities of updates being read, we can construct a pool with a specified probability mass cover.

FUTURE WORK: Test-collections based on probability-based pooling; Are they effective for measuring relative system performance? Are they reusable? How might they work for different user models?

ACKNOWLEDGEMENT: Support provided in part by an ICTIR 2015 Student Travel Grant.