Machine Assisted Rapid Review

Max Callaghan, Finn Müller-Hansen

MCC

July 3, 2019

Screening

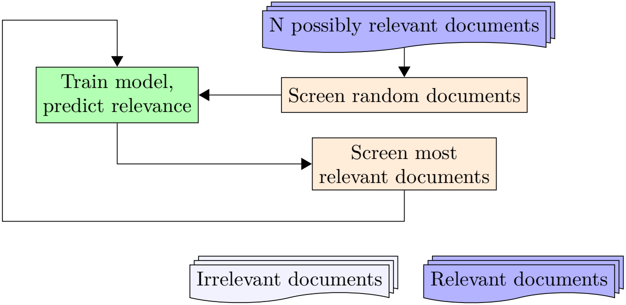
In systematic review, we want to have a broad search query, to make sure we include as many relevant documents as possible

This means we need to screen all of the documents the search query returns, in order to find a clean set of relevant documents.

Machine Assisted Screening

- Scientific literature is expanding in all subject areas, increasing the work needed to conduct systematic reviews.
- The process generates training data, making it an ideal setting for applying *Active Learning* see Settles (2009)
- An expanding literature exists on using active learning for screening in systematic reviews O'Mara-Eves et al. (2015)

Active Learning



When do we stop? How can we describe when we stopped, and what we think this means about the literature we have selected and not selected?
$Precision = 1.0$, but what about $Recall$?

Missing stopping criteria

Research on a Stopping criterion in AL literature suggest stopping when the confidence or performance of the classifier drops. These indicate when the ML part is no longer generating useful results, but do not relate to recall: Vlachos stress that "the remaining instances should be considered redundant only for the given model and feature representation" Vlachos (2008)

Stopping criteria in the systematic review literature are not well articulated:

> *Once a given stopping criterion is reached (for example, when all relevant studies have been identified, or when the reviewers have run out of time for manual screening), the process ceases, and the remainder of studies not yet screened manually is discarded Miwa et al. (2014)*

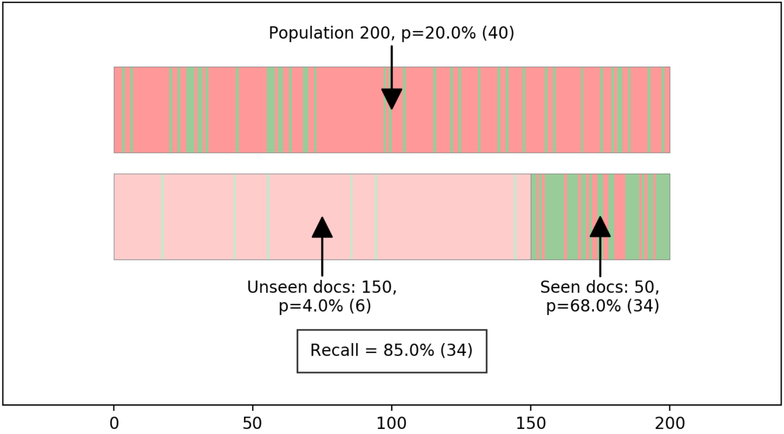
Missing stopping criteria

Statistical approaches are hinted at but underdeveloped:

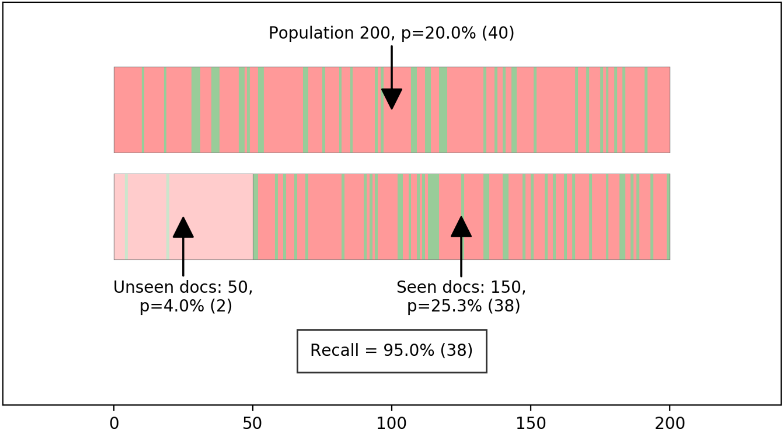
> *Deciding when to stop can be left to the end-user, eg, if the prioritised references are consistently irrelevant, or heuristics or statistical approaches (based on an unbiased sample of the remaining references) can inform this decision. We intend to investigate this problem in the future Przybyła et al. (2018)*

A reliable stopping criterion is still a research gap:
"The problem of choosing a cut-off threshold, equivalent to deciding when to stop when using a model for prioritising relevant documents, remains an open research question in information retrieval" Bannach-Brown et al. (2019)

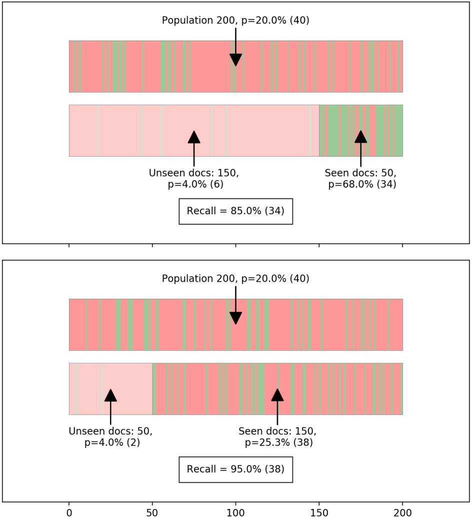
Relevance of unseen documents is insufficient



Relevance of unseen documents is insufficient

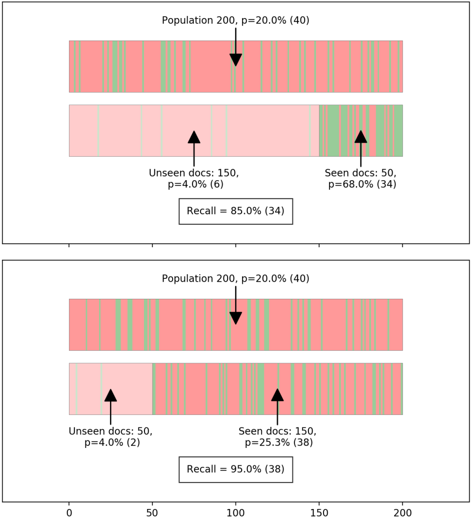


Relevance of unseen documents is insufficient



It doesn't just matter what proportion of unseen documents is relevant, but also what proportion of documents is unseen

Relevance of unseen documents is insufficient



It doesn't just matter what proportion of unseen documents is relevant, but also what proportion of documents is unseen

We can calculate the recall with the relevance of the unseen docs, the relevance of the population and porportion of documents seen

$$1 - \frac{pU}{pP}\frac{U}{P}$$

Uncertainty of pU

However, without perfect knowledge, we don't *know* the relevance of unseen documents: but we can *estimate* it through taking a sample. We do know U, the number of unseen documents, S, the number of seen documents, and pS, the relevance of seen documents,

For a given p-value, we can generate an Agresti-Coull interval

$$CI_{AC} = \tilde{p} \pm \kappa(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2}$$

We can use the outer limit $\dot{p} = \max \tilde{p}$ to estimate the relevance of all documents

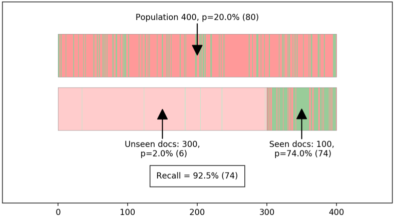
$$\dot{p}P = \frac{pS*S+\dot{p}U*U}{N}$$

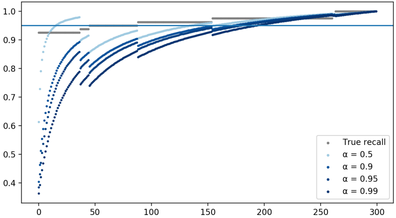
And plug both of these into the original equation to generate a minimum estimated recall

$$\hat{R} = 1 - \frac{\dot{p}U}{\dot{p}P}\frac{U}{P}$$

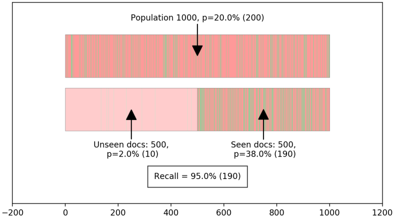
We can set the recall and confidence interval prior to the start of screening

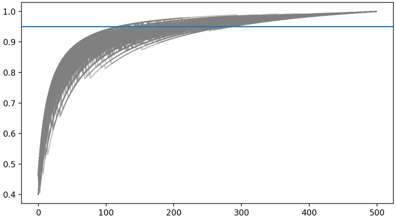
Uncertainty of pU


Population 400, p=20.0% (80)
Unseen docs: 300, p=2.0% (6)
Seen docs: 100, p=74.0% (74)
Recall = 92.5% (74)
0
100
200
300
400



Uncertainty of pU

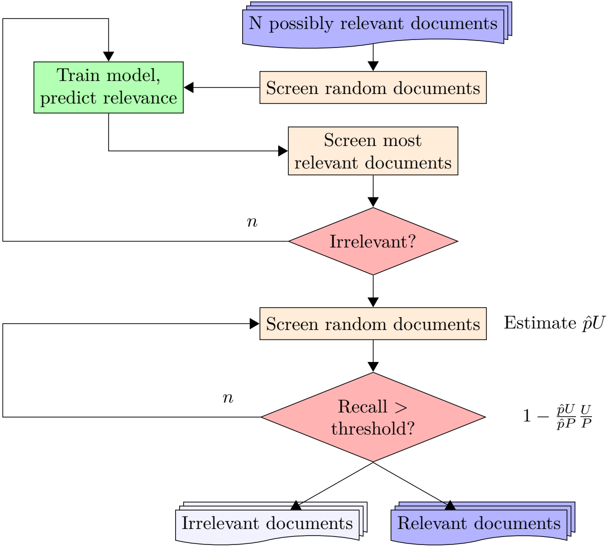




$$CI_{AC} = \tilde{p} \pm \kappa(\tilde{p}\tilde{q})^{1/2}\tilde{n}^{-1/2}$$

$$\hat{R} = 1 - \frac{\dot{p}U}{\dot{p}P}\frac{U}{P}$$

Active Learning with Robust Probabilistic Stopping Criteria



N possibly relevant documents

Train model, predict relevance

Screen random documents

Screen most relevant documents

Irrelevant?

n

Screen random documents

Estimate $\hat{p}U$

Recall > threshold?

n

$1 - \frac{\hat{p}U}{\hat{p}P}\frac{U}{P}$

Irrelevant documents

Relevant documents

Evaluation on existing dataset(s)

To be done - ours? Someone elses? Simulations with many types/specifications of ML models, showing reduction in work, predicted and actual recall

Necessary?

How big should the sample be?

What confidence intervals do we get for different levels of relevance and different sample sizes?

Increasing N reduces unseen docs and size of confidence interval

Do we need to take a sample?

Could we just keep looking at the most relevant documents (this would make our estimates more conservative) or does this just make things complicated without saving time?

When do we move from the most relevant documents to a random sample?

Is this when we find no more relevant documents? When no more relevant documents are predicted? When model performance on new data drops below a certain level? Some other criterion or combination of criteria?

Do we need to rate all documents in the sample?

Can the system estimate recall as we rate documents, and stop as soon as we have reached the threshold?

What if we prioritised irrelevant documents?

In this way we could be certain about recall, and uncertain about precision.

We spend a lot of time looking at irrelevant documents (so learn less about the literature during the process), but sometimes exclusion decisions can be faster than inclusion decisions.

Conclusion

This is a new way to use machine learning in screening

- Offers robust stopping criteria

Conclusion

This is a new way to use machine learning in screening

- Offers robust stopping criteria
- ML model performance does not determine accuracy of document selection, only influences work reduction

References

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S. C., Ananiadou, S., Liao, J., and Macleod, M. R. (2019). Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, 8(1):1–12.

Miwa, M., Thomas, J., O'Mara-Eves, A., and Ananiadou, S. (2014). Reducing systematic review workload through certainty-based screening. *Journal of Biomedical Informatics*, 51:242–253.

O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., and Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1):1–22.

Przybyła, P., Brockmeier, A. J., Kontonatsios, G., Le Pogam, M. A., McNaught, J., von Elm, E., Nolan, K., and Ananiadou, S. (2018). Prioritising references for systematic reviews with RobotAnalyst: A user study. *Research Synthesis Methods*, 9(3):470–488.

Settles, B. (2009). Active Learning Literature Survey. Technical report, University of Wisonsin-Madison.

Vlachos, A. (2008). A stopping criterion for active learning. *Computer Speech and Language*, 22(3):295–312.