Partially automated literature screening for systematic reviews by modelling non-relevant articles

Henry Petersen[1] and Josiah Poon[1] Simon Poon[1] Clement Loy[2] Mariska Leeflang[3]

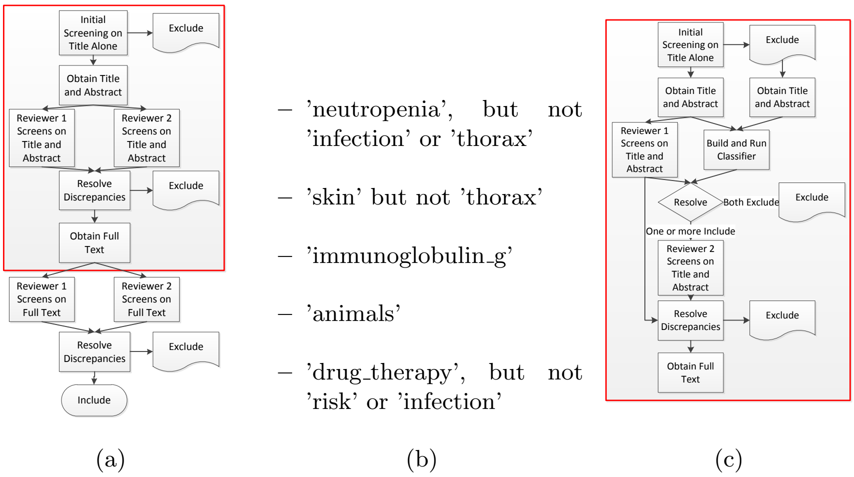
[1] School of Information Technologies, University of Sydney, Australia
[2] School of Public Health, University of Sydney, Australia
[3] Academic Medical Center, University of Amsterdam, Netherlands
hpet9515@uni.sydney.edu.au, {josiah.poon,simon.poon}@sydney.edu.au,
clement.loy@sydney.edu.au, m.m.leeflang@amc.uva.nl

Systematic reviews are widely considered as the highest form of medical evidence, since they aim to be a repeatable, comprehensive, and unbiased summary of the existing literature. Because of the high cost of missing relevant studies, review authors go to great lengths to ensure all relevant literature is included. It is not atypical for a single review to be conducted over the course of months or years, with multiple authors screening thousands of articles in a multi-stage triage process; first on title, then on title and abstract, and finally on full text. Figure 1a shows a typical literature screening process for systematic reviews.

In the last decade, the information retrieval (IR) and machine learning (ML) communities have shown increasing interest in literature searches for systematic reviews [1–3]. Literature screening for systematic reviews can be characterised as a classification task with two defining features; a requirement for near perfect recall on the class of relevant studies (the high cost of missing relevant evidence), and highly imbalanced training data (review authors are often willing to screen thousands of citations to find less than 100 relevant articles). Previous attempts at automating literature screening for systematic reviews have primarily focused on two questions; how to build a suitably high recall model for the target class in a given review under the conditions of highly imbalanced training data [1, 3], and how best to integrate classification into the literature screening process [2].

When screening articles, reviewers exclude studies for a number of reasons (animal populations, incorrect disease etc.). Additionally, in any given triage stage a study may not be relevant but still progress to the next stage as the authors have insufficient information to exclude it (i.e. the title may not indicate a study was performed with an animal population, however this may become apparent upon reading the abstract). We meet the requirement for near perfect recall on relevant studies by inverting the classification task and identifying subsets of irrelevant studies with near perfect precision. We attempt to identify such studies by training the classifier using the labels assigned at the previous triage stage (see Figure 1c). The seamless integration with the existing manual screening process is an advantage of our approach.

The classifier is built by first selecting terms from the title and abstracts with the greatest information gain on labels assigned in the first triage stage. Articles


Initial Screening on Title Alone
Exclude
Obtain Title and Abstract
Reviewer 1 Screens on Title and Abstract
Reviewer 2 Screens on Title and Abstract
Resolve Discrepancies
Exclude
Obtain Full Text
Reviewer 1 Screens on Full Text
Reviewer 2 Screens on Full Text
Resolve Discrepancies
Exclude
Include
(a)
(b)
Initial Screening on Title Alone
Exclude
Obtain Title and Abstract
Obtain Title and Abstract
Reviewer 1 Screens on Title and Abstract
Build and Run Classifier
Resolve
Both Exclude
Exclude
One or more Include
Reviewer 2 Screens on Title and Abstract
Resolve Discrepancies
Exclude
Obtain Full Text
(c)

- 'neutropenia', but not 'infection' or 'thorax'
- 'skin' but not 'thorax'
- 'immunoglobulin_g'
- 'animals'
- 'drug_therapy', but not 'risk' or 'infection'

Fig. 1: Typical literature screening process for systematic reviews, sample rules generated by our classifier, and the proposed modified screening process.

are then represented as Boolean statements over these terms, and interpretable rules are then generated using Boolean minimisation (examples of rules are given in 1b Review authors can then refine the classifier by selecting only those rules most likely to describe non-relevant studies, maximising overall precision.

Preliminary experiments simulating the process outlined in Figure 1c on a previously conducted systematic review indicate that as many as 25% of articles can be safely eliminated without the need for screening by a second reviewer. The evaluation does assume that all false positives (studies erroneously excluded by the generated rules) were included by the first reviewer. Such an assumption is reasonable; the reason for multiple reviewers is that even human experts make mistakes. A study comparing the precision of our classifier to human reviewers is planned. In addition, future work will focus on improving the quality of the generated rules by trying to better capture reasons for excluding studies matching those used by human reviewers.

References

1. Aaron M. Cohen, Kyle H. Ambert, and Marian McDonagh. Research paper: Cross-topic learning for work prioritization in systematic review creation and update. *JAMIA*, 16(5):690–704, 2009.
2. Oana Frunza, Diana Inkpen, Stan Matwin, William Klement, and Peter OBlenis. Exploiting the systematic review protocol for classification of medical abstracts. *Artificial Intelligence in Medicine*, 51(1):17 – 25, 2011.
3. Stan Matwin, Alexandre Kouznetsov, Diana Inkpen, Oana Frunza, and Peter O'Blenis. A new algorithm for reducing the workload of experts in performing systematic reviews. *JAMIA*, 17(4):446–453, 2010.