Automated Forward Citation Snowballing using Google Scholar and Machine Learning

Hung-Yuan (Vincent) Cheng [1, 2]

[1] Department of Population Health Sciences, Bristol Medical School, University of Bristol, UK
[2] NIHR Bristol Biomedical Research Centre, Bristol, UK

NHS
University Hospitals Bristol
NHS Foundation Trust

Motivation

- A systematic review summarises the results of available studies and provides a high level of evidence on the effectiveness of healthcare interventions to inform recommendations for healthcare.
- In a systematic review, searching for studies is one of the most crucial steps. Forward citation snowballing (refers to identify new studies based on those papers citing the study being examined) is an effective method to look for new studies and double-check.
- Google Scholar is a comprehensive database for forward citation snowballing.

Problem

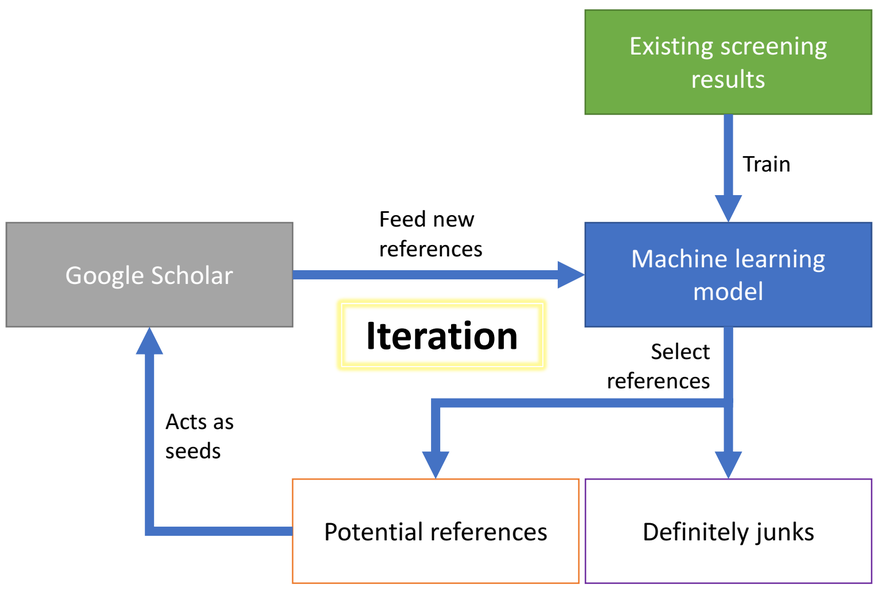
- Google Scholar's searching algorithm and display are not reproducible and transparent.
- The search results can come up with many irrelevant studies, leading to labour- and time-consuming screening.
- Google Scholar does not provide an easy interface for downloading a large amount of references.

Aims

1. To observe Google Scholar searching algorithm and citation pattern for forward citation snowballing
2. To build an automated Google Scholar scraping system
3. To filter out irrelevant studies by machine learning

System overview

- An automated forward citation snowballing system using Google Scholar search and machine learning



Reflection

- Duplicates occurred more frequently in later iterations, reflecting decreased true retrieval rate and inefficient retrievals.
- After manually screened the retrieved references, useful references appeared more frequently in the first 4 iterations.
- Machine learning model still has room to improve its performance.

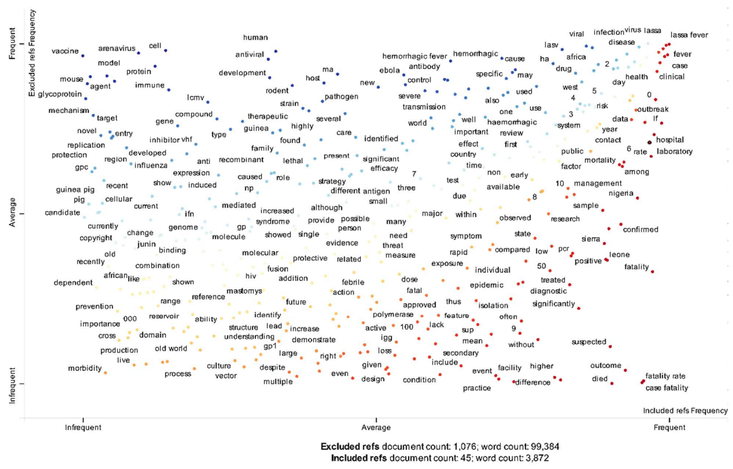
Future work

- Automated network plots and other visualisation features
- Tests on a variety of topics and parameters to determine optimal performance
- Other searching databases, such as Microsoft Academy, CrossRef, and Semantic Scholar

Implementation

Machine learning model

- Existing screening results: excluded (n=1076) vs included references (n=45)
- Four separate machine learning models trained by individual inputs: 1) authors; 2) title; 3) journal and 4) abstract from each reference
- Text data processing and cleaning [nltk]
- Data exploration [scattertext]
- Machine learning (ML) model training via *Tokenization*, *Tf-IDF*, and *RandomForestClassifier* [sklearn]
- Accuracy ~ 86% in four models
- Four ML models form a majority rule system to exclude irrelevant references



Scan me

Google Scholar scraper

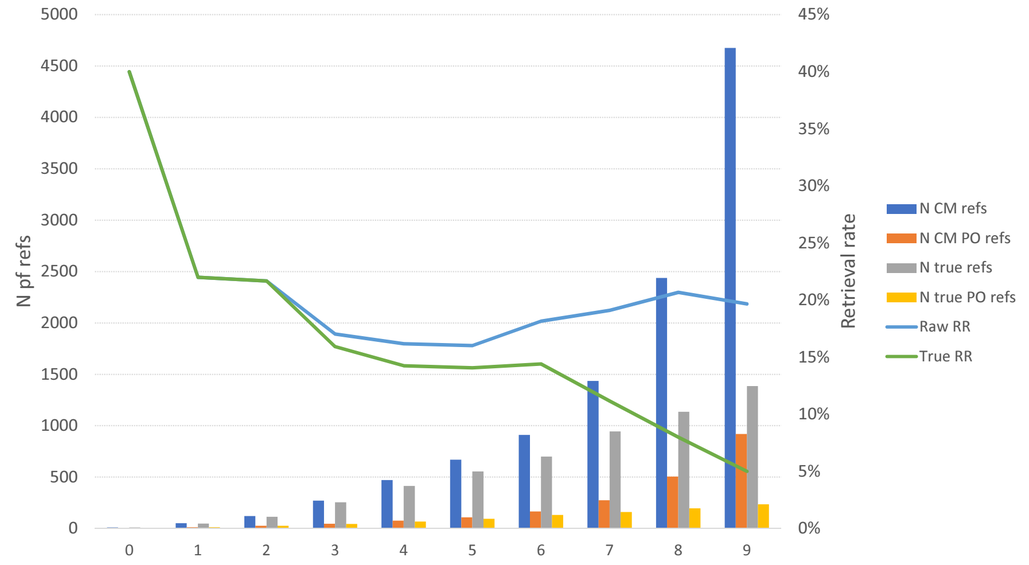
- Scrape key data from Google Scholar search results, including Authors, Journal, Title, Abstract, Publication Year, URL, URL of cited, Number of cited [selenium/request-html/urllib]

Automated forward citation snowballing system

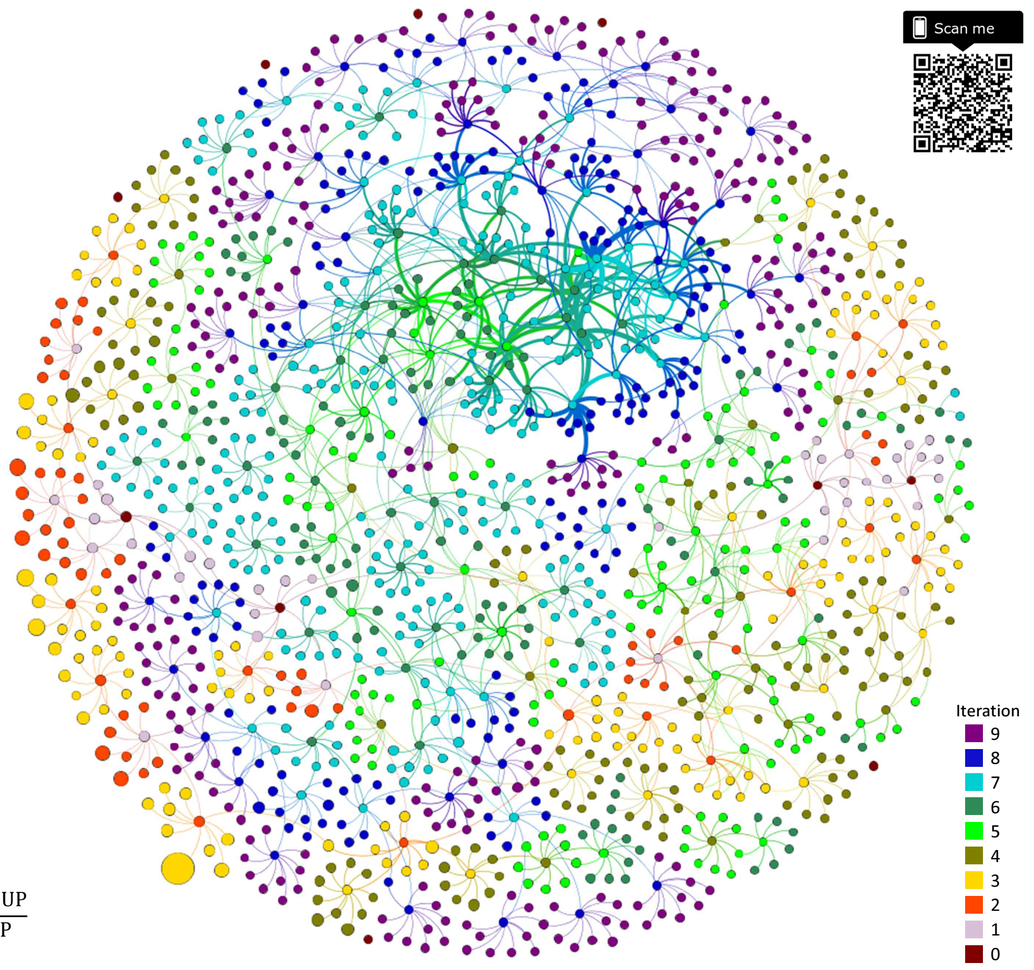
- Started with key words, "Lassa fever ribavirin"
- Retrieved up to 10 references from "cited by" link in each potential reference
- Ten iterations:

IT	N CM refs	N CM PO refs
0	10	4
1	50	11
2	120	26
3	270	46
4	470	76
5	668	107
6	909	165
7	1434	274
8	2437	504
9	4675	919

IT	N CM ref (de-DUP)	N CM PO refs (de-DUP)
0	10	4
1	47	11
2	114	26
3	254	43
4	414	67
5	555	94
6	699	131
7	944	160
8	1134	195
9	1384	234



Scan me



Glossary
CM: cumulative
DUP: duplicate
IT: iteration
N: Number
PO: potential
RR: Retrieval rate

Metric

$$\text{Raw RR} = \frac{\text{N of CM PO refs}}{\text{N of CM refs}}$$

$$\text{True RR} = \frac{\text{N of CM PO refs after removing DUP}}{\text{N of CM refs after removing DUP}}$$

Acknowledgement

Thanks to the Jean Golding Institute, University of Bristol for their generous support. Also, thanks to authors behind open source code for contribution and Auto-synthesis group, University of Bristol for inspiration.

Funded by NIHR