Testing a Citation and Text-Based Framework for Retrieving Publications for Literature Reviews

M. Janina Sarol (mjsarol@illinois.edu)

Linxi Liu (lliu73@illinois.edu)

Jodi Schneider (jodi@illinois.edu)

ILLINOIS
School of Information Sciences

Literature Reviews

- Scoping review
- State-of-the-art review
- Rapid review
- Systematic review

Grant, M.J., Booth, A.: A typology of reviews: an analysis of 14 review types and associated methodologies. Health Information & Libraries Journal **26**(2) (2009) 91–108

Problem with Current Process

- Hard to find and identify relevant papers
  - Growing literature base

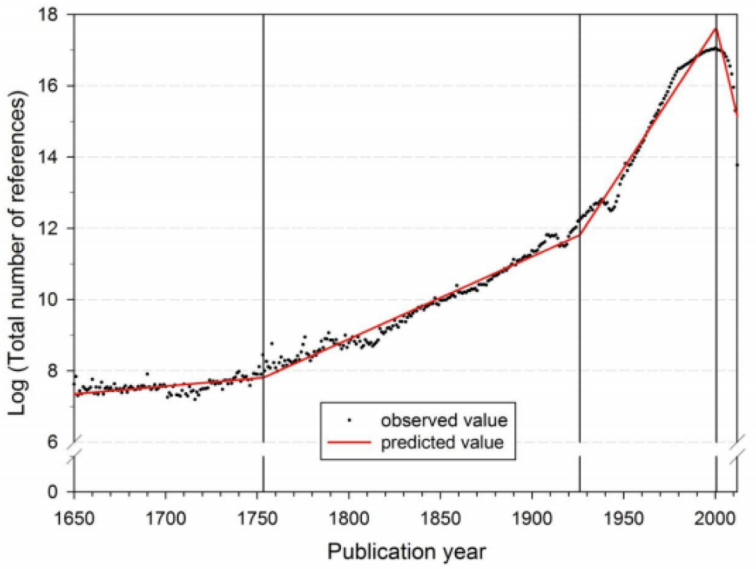
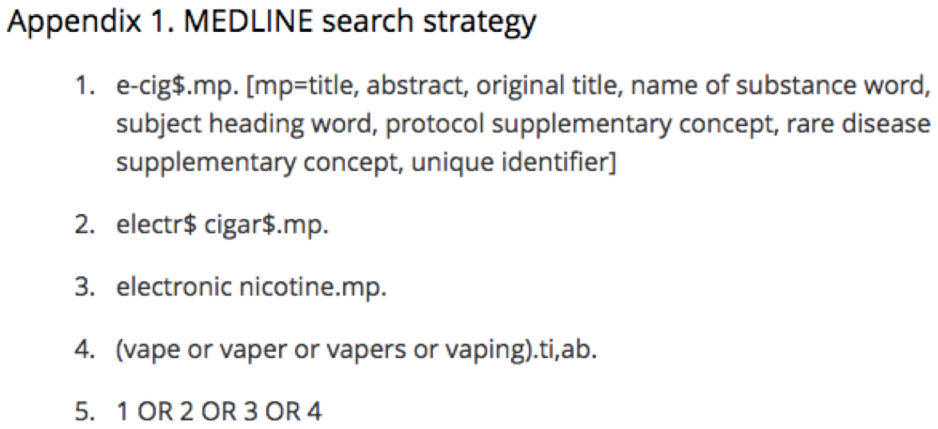


Figure 2. Segmented growth of the annual number of cited references from 1650 to 2012 (citing publications from 1980 to 2012)

Image from: Bornmann, L., Mutz, R.: Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. Journal of the Associa-tion for Information Science and Technology 66 (11) (2015) 2215–2222

Problem with Current Process

- Hard to find and identify relevant papers
  - Growing literature base
  - Time-consuming
    - To develop a search protocol



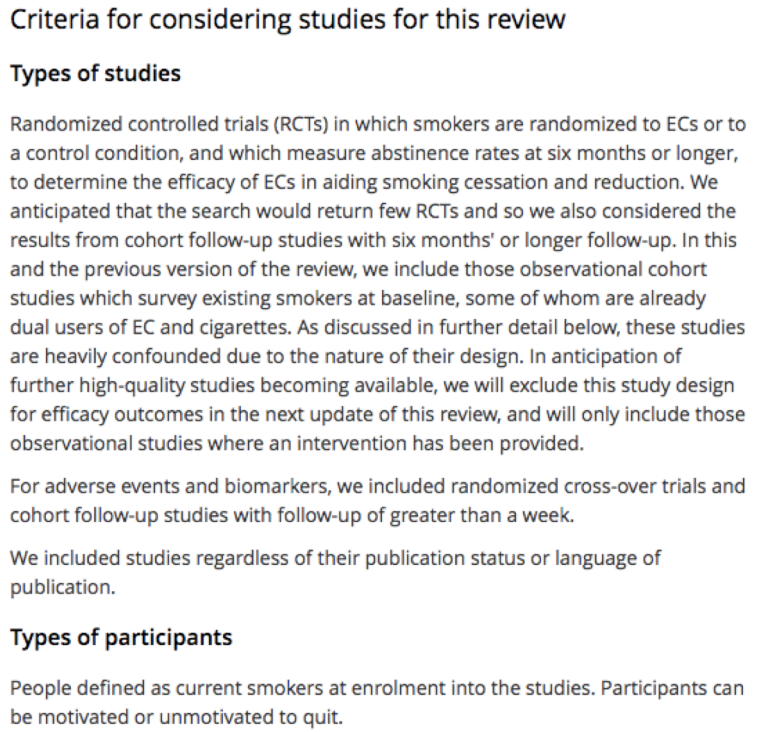
Appendix 1. MEDLINE search strategy

1. e-cig$.mp. [mp=title, abstract, original title, name of substance word, subject heading word, protocol supplementary concept, rare disease supplementary concept, unique identifier]
2. electr$ cigar$.mp.
3. electronic nicotine.mp.
4. (vape or vaper or vapers or vaping).ti,ab.
5. 1 OR 2 OR 3 OR 4

Screenshot from: McRobbie, H., Bullen, C., Hartmann-Boyce, J., Hajek, P.: Electronic cigarettes or smoking cessation and reduction. Cochrane Database of Systematic Reviews 12 (2014)

Problem with Current Process

- Hard to find and identify relevant papers
  - Growing literature base
  - Time-consuming
    - To develop a search protocol
    - To screen papers for relevance



Criteria for considering studies for this review

Types of studies

Randomized controlled trials (RCTs) in which smokers are randomized to ECs or to a control condition, and which measure abstinence rates at six months or longer, to determine the efficacy of ECs in aiding smoking cessation and reduction. We anticipated that the search would return few RCTs and so we also considered the results from cohort follow-up studies with six months' or longer follow-up. In this and the previous version of the review, we include those observational cohort studies which survey existing smokers at baseline, some of whom are already dual users of EC and cigarettes. As discussed in further detail below, these studies are heavily confounded due to the nature of their design. In anticipation of further high-quality studies becoming available, we will exclude this study design for efficacy outcomes in the next update of this review, and will only include those observational studies where an intervention has been provided.

For adverse events and biomarkers, we included randomized cross-over trials and cohort follow-up studies with follow-up of greater than a week.

We included studies regardless of their publication status or language of publication.

Types of participants

People defined as current smokers at enrolment into the studies. Participants can be motivated or unmotivated to quit.

Screenshot from: McRobbie, H., Bullen, C., Hartmann-Boyce, J., Hajek, P.: Electronic cigarettes or smoking cessation and reduction. Cochrane Database of Systematic Reviews 12 (2014)

Problem with Current Process

- Hard to find and identify relevant papers
  - Growing literature base
  - Time-consuming
    - To develop a search protocol
    - To screen papers for relevance
- Manual approach

How can we solve this problem?

- Skip the development of the search strings
- Develop new search method
  - Achieve higher precision (without affecting recall)
  - Retrieve fewer results -> screen fewer documents
- Automate the screening process

Start the search from 1 (or more) paper(s) to be included in the literature review

seed papers

How do we find the other papers using the seed paper(s)?

Use both text and citation data

- The papers to be included in the literature review have some (topical) similarity
- Both text and citation data have been used to model the similarity between papers
  - More shared terms = higher similarity
  - More connections in citation network = higher similarity

Proposed Framework

- **Select** a set of seed paper(s)
- **Search**: collect papers connected through various citation relationships
- **Filter**: keep papers similar to the seed paper(s)
  - Citation-based
  - Text-based

Sample Implementation and Experiment

#	Article Title	Manually Reviewed	Included Studies
1	Antibiotic regimens for management of intra-amniotic infection	1,001	11
2	Interventions for preventing and ameliorating cognitive deficits in adults treated with cranial irradiation	2,762	6
3	Co-enzyme Q10 supplementation for the primary prevention of cardiovascular disease	1,348	6
4	Intermittent self-dilatation for urethral stricture disease in males	276	11
5	Electronic cigarettes for smoking cessation and reduction	594	13
6	Long-term proton pump inhibitor (PPI) use and the development of gastric pre-malignant lesions	502	8

List of reviews taken from: Belter, C.W.: Citation analysis as a literature search method for systematic reviews. Journal of the Association for Information Science and Technology 67 (11) (2016) 2766–2777

Sample Implementation and Experiment

#	Article Title	Manually Reviewed	Included Studies
1	Antibiotic regimens for management of intra-amniotic infection	1,001	11
2	Interventions for preventing and ameliorating cognitive deficits in adults treated with cranial irradiation	2,762	6
3	Co-enzyme Q10 supplementation for the primary prevention of cardiovascular disease	1,348	6
4	Intermittent self-dilatation for urethral stricture disease in males	276	11
5	Electronic cigarettes for smoking cessation and reduction	594	13
6	Long-term proton pump inhibitor (PPI) use and the development of gastric pre-malignant lesions	502	8

GOAL: retrieve all studies

List of reviews taken from: Belter, C.W.: Citation analysis as a literature search method for systematic reviews. Journal of the Association for Information Science and Technology 67 (11) (2016) 2766–2777

Sample Implementation and Experiment

- Data source: Scopus
- Data collection:
  - Scopus Search API
  - Scopus Abstract Retrieval API
- Only used major publications for the studies
- Mimic initial search conditions (based on year)

Electronic searches

We searched the following databases in July 2014:

- Cochrane Tobacco Addiction Group Specialised Register
- Cochrane Central Register of Controlled Trials (CENTRAL) (*The Cochrane Library* , Issue 7, 2014)

Screenshot from: McRobbie, H., Bullen, C., Hartmann-Boyce, J., Hajek, P.: Electronic cigarettes or smoking cessation and reduction. Cochrane Database of Systematic Reviews 12 (2014)

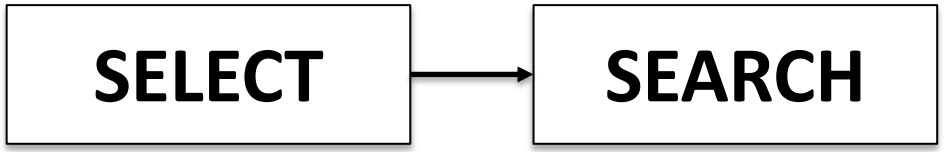
Sample Implementation and Experiment

SELECT

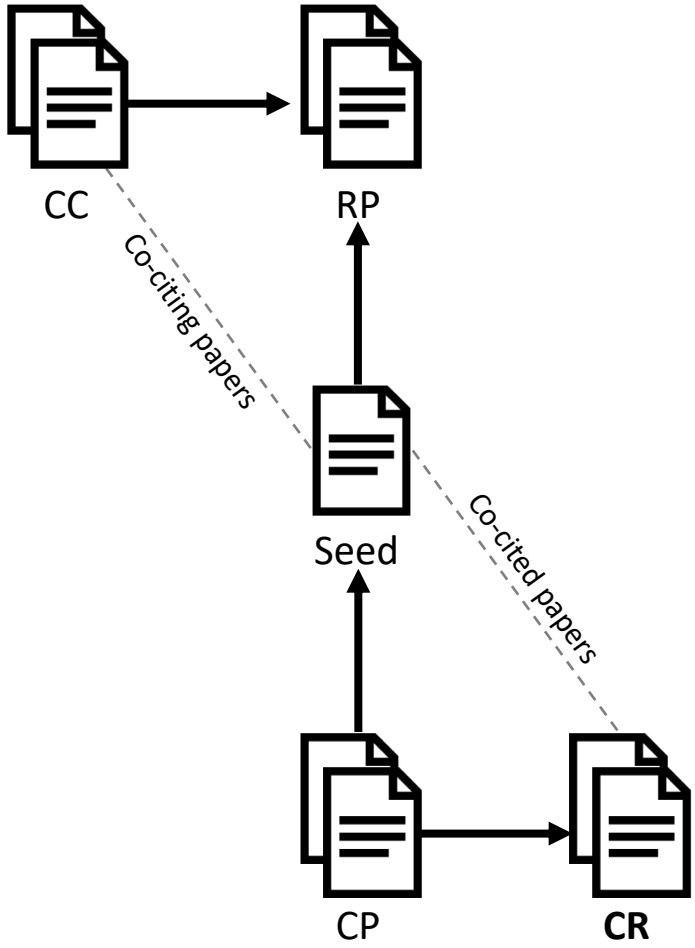
- 1 seed
- All combinations of 2 seeds – 55 combinations
- All combinations of 3 seeds – 165 combinations

#	Article Title	Manually Reviewed	Included Studies
1	Antibiotic regimens for management of intra-amniotic infection	1,001	11

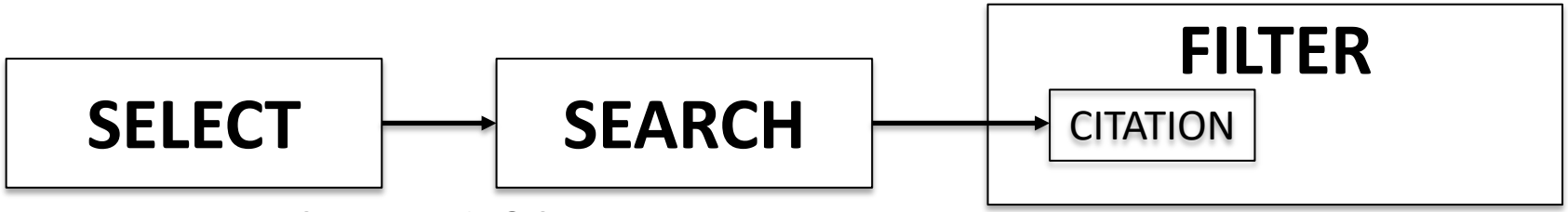
Sample Implementation and Experiment



- References (RP)
- Citations (CP)
- Co-citing papers (CC)
- Co-cited papers (CR)

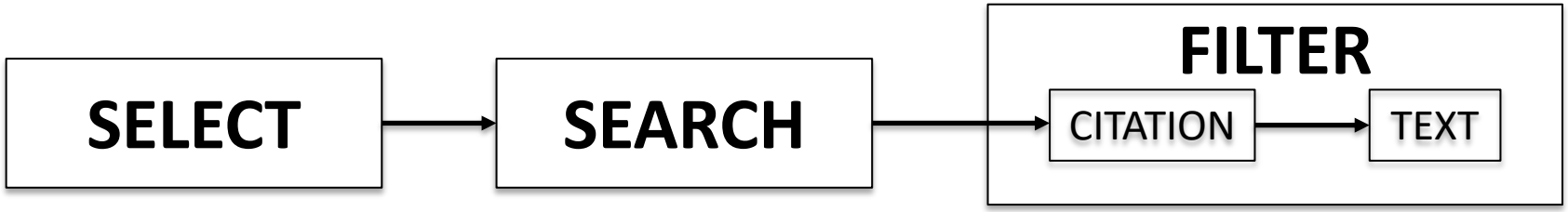


Sample Implementation and Experiment



- Citation-based filtering
  - Reference: seed paper cites paper A
  - Citation: seed paper is cited by paper A
  - Co-citing: seed paper shares at least 10% of its references with paper A
  - Co-cited: seed paper shares at least 10% of its citations with paper A
- If any of the above is true, keep paper A!

Sample Implementation and Experiment



- Text-Based Filtering
  - Extracted keywords from the abstracts
    - Extraction tool: Rapid Automatic Keyword Extraction
    - Used only the **bigram** and **trigram** keywords
  - Keep paper A if there is at least **1 keyword match** with the seed paper

Experiment Results: 1 seed

Review	Manually Reviewed	Documents Retrieved (Avg)	Included Studies	In Scopus	Avg	Max	Min
1	1,001	86	11	10	4	6	1
2	2,762	84	6	4	1.25	2	1
3	1,348	146	6	5	2	3	1
4	276	45	11	11	4.45	8	1
5	594	75	13	10	4	7	1
6	502	140	8	8	3.38	5	3

Experiment Results: 1 seed

Review	Manually Reviewed	Documents Retrieved (Avg)	Included Studies	In Scopus	Avg	Max	Min
1	1,001	86	11	10	4	6	1
2	2,762	84	6	4	1.25	2	1
3	1,348	146	6	5	2	3	1
4	276	45	11	11	4.45	8	1
5	594	75	13	10	4	7	1
6	502	140	8	8	3.38	5	3

Not all studies are in Scopus

Experiment Results: 1 seed

Review	Manually Reviewed	Documents Retrieved (Avg)		Included Studies	In Scopus	Avg	Max	Min
1	1,001	86	-915	11	10	4	6	1
2	2,762	84	-2678	6	4	1.25	2	1
3	1,348	146	-1202	6	5	2	3	1
4	276	45	-231	11	11	4.45	8	1
5	594	75	-519	13	10	4	7	1
6	502	140	-362	8	8	3.38	5	3

Higher precision

Experiment Results: 1 seed

Review	Manually Reviewed	Documents Retrieved (Avg)	Included Studies	In Scopus	Avg	Max	Min
1	1,001	86	11	10	4	6	1
2	2,762	84	6	4	1.25	2	1
3	1,348	146	6	5	2	3	1
4	276	45	11	11	4.45	8	1
5	594	75	13	10	4	7	1
6	502	140	8	8	3.38	5	3

Average relevant results retrieved (all combinations)

Experiment Results: 1 seed

Review	Manually Reviewed	Documents Retrieved (Avg)	Included Studies	In Scopus	Avg	Max	Min
1	1,001	86	11	10	4	6	1
2	2,762	84	6	4	1.25	2	1
3	1,348	146	6	5	2	3	1
4	276	45	11	11	4.45	8	1
5	594	75	13	10	4	7	1
6	502	140	8	8	3.38	5	3

# of results retrieved by best combination

Experiment Results: 1 seed

Review	Manually Reviewed	Documents Retrieved (Avg)	Included Studies	In Scopus	Avg	Max	Min
1	1,001	86	11	10	4	6	1
2	2,762	84	6	4	1.25	2	1
3	1,348	146	6	5	2	3	1
4	276	45	11	11	4.45	8	1
5	594	75	13	10	4	7	1
6	502	140	8	8	3.38	5	3

# of results retrieved by worst combination

Experiment Results: 2 seeds

Review	Manually Reviewed	Documents Retrieved (Avg)	Included Studies	In Scopus	Avg	Max	Min
1	1,001	223	11	10	6.91	9	4
2	2,762	210	6	4	2.83	**4**	2
3	1,348	342	6	5	3.7	**5**	2
4	276	113	11	11	7.13	9	4
5	594	163	13	10	6.36	9	3
6	502	332	8	8	5.46	**8**	3

Experiment Results: 3 seeds

Review	Manually Reviewed	Documents Retrieved (Avg)	Included Studies	In Scopus	Avg	Max	Min
1	1,001	381	11	10	8.4	**10**	6
2	2,762	343	6	4	3.75	**4**	3
3	1,348	542	6	5	4.5	**5**	4
4	276	181	11	11	8.4	**11**	6
5	594	250	13	10	7.72	**10**	6
6	502	530	8	8	6.75	**8**	4

Advantages of Framework

- Fewer results
  - Fewer documents to be read and screened
- Automation
  - Can use existing APIs to collect data
- Potential for more coverage
  - Conventional search methods might miss papers

Limitations

- Relies on the comprehensive coverage of the source database(s)
  - Only 48 out of 55 included studies in our experiments were in Scopus
  - Missing documents:
    - 6 journal articles
    - 1 conference abstract
  - Incomplete information:
    - 1 no abstract
    - 9 no list of references

Future Work

- Test the framework for 7,158 systematic reviews
- Conduct a detailed analysis of seed selection
  - Which types of seeds perform best?
- Use framework in on-going literature reviews
  - How can this framework be integrated in the current literature review processes?
- Other implementations of the framework
  - Using topic modeling