NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

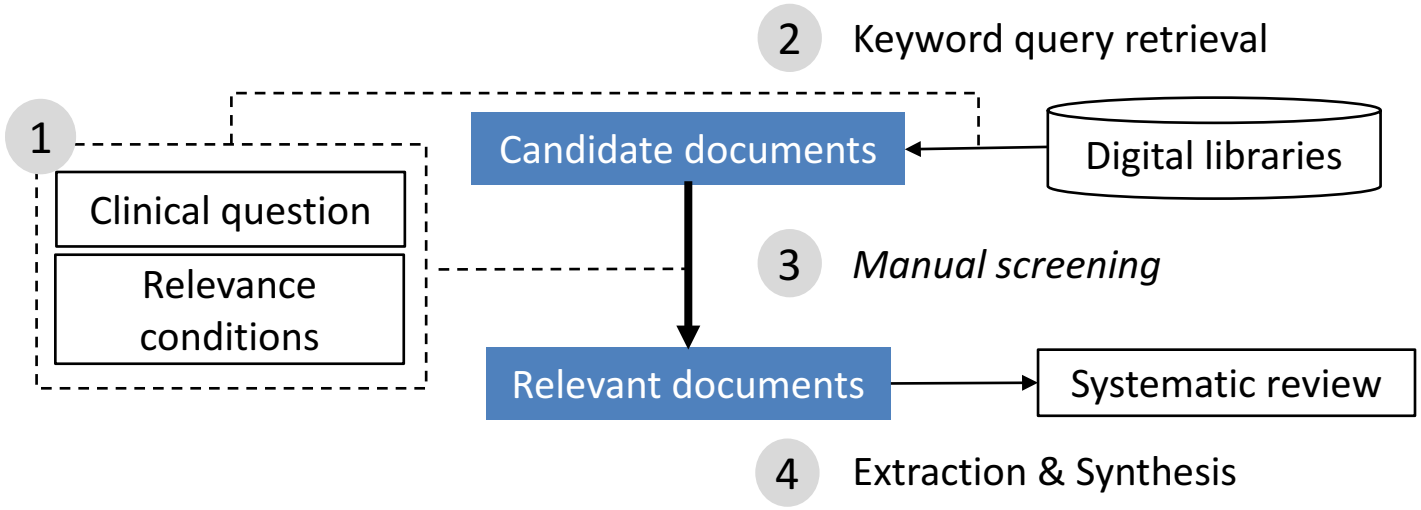
Seed-driven Document Ranking for Systematic Reviews in Evidence-Based Medicine

Grace E. Lee and Aixin Sun
School of Computer Science and Engineering
Nanyang Technological University (NTU), Singapore

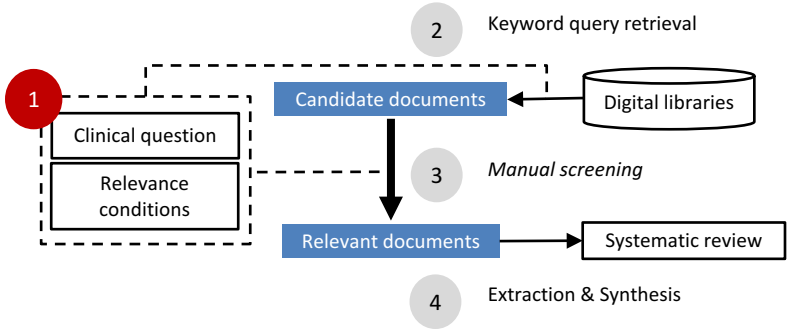
Systematic Reviews

- Literature survey
  - providing conclusions of clinical questions (topics)
  - existing literature
  - state-of-the-art answer of the clinical question

- SRs are conducted by following systematic steps



1. Defining a Clinical Question

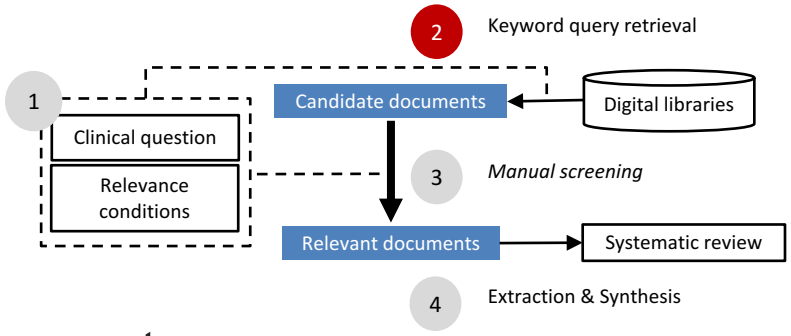


- Set up a clinical question (topic)
  - existing biomedical literature
  - one or two relevant publications

- Define relevance conditions (eligibility criteria)
  - evaluating relevance of documents
  - explicit details

Patient	Intervention	Comparator	Outcome	Study type
• pancreatic cancer • seniors (>60) • surgical and medical history	• laparoscopy • laparotomy • endoscopy	• physical examination • surgical examination	• staging of cancer cell • resectability of cancer cell	• randomized controlled test • comparative study • prospective study

2. Retrieval Process

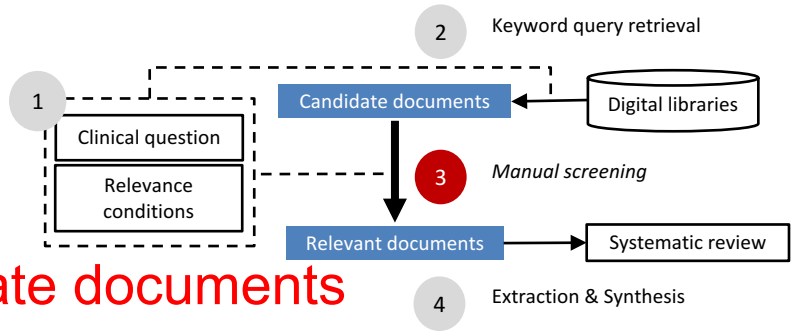


- Collecting candidate documents
  - without missing out any relevant documents
  - high recall

- Various keyword queries to multiple databases
  - PubMed, MEDLINE, EMBASE, Cochrane CENTRAL

- Large candidate collection
  - more than 2,000 candidate documents in general for one SR

3. Screening Process

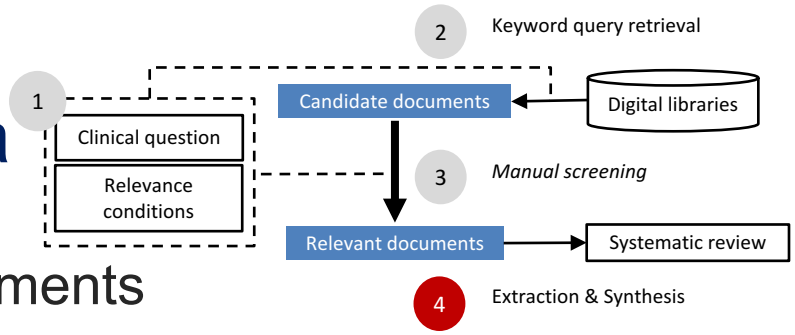


- Identify relevant documents in candidate documents
  - manual screening
  - multiple SR experts
  - detailed relevance conditions

- Output of screening step
  - relevant documents
  - 1 to 2 percent of candidate documents

Sample SRs	#Rel Docs	#Candidate Docs	% of Rel Docs
SR_1	16	1,911	0.83%
SR_2	48	10,872	0.44%
SR_3	3	1,573	0.19%
SR_4	6	2,065	0.29%
SR_5	46	5,971	0.77%

4. Extract and Synthesize Data

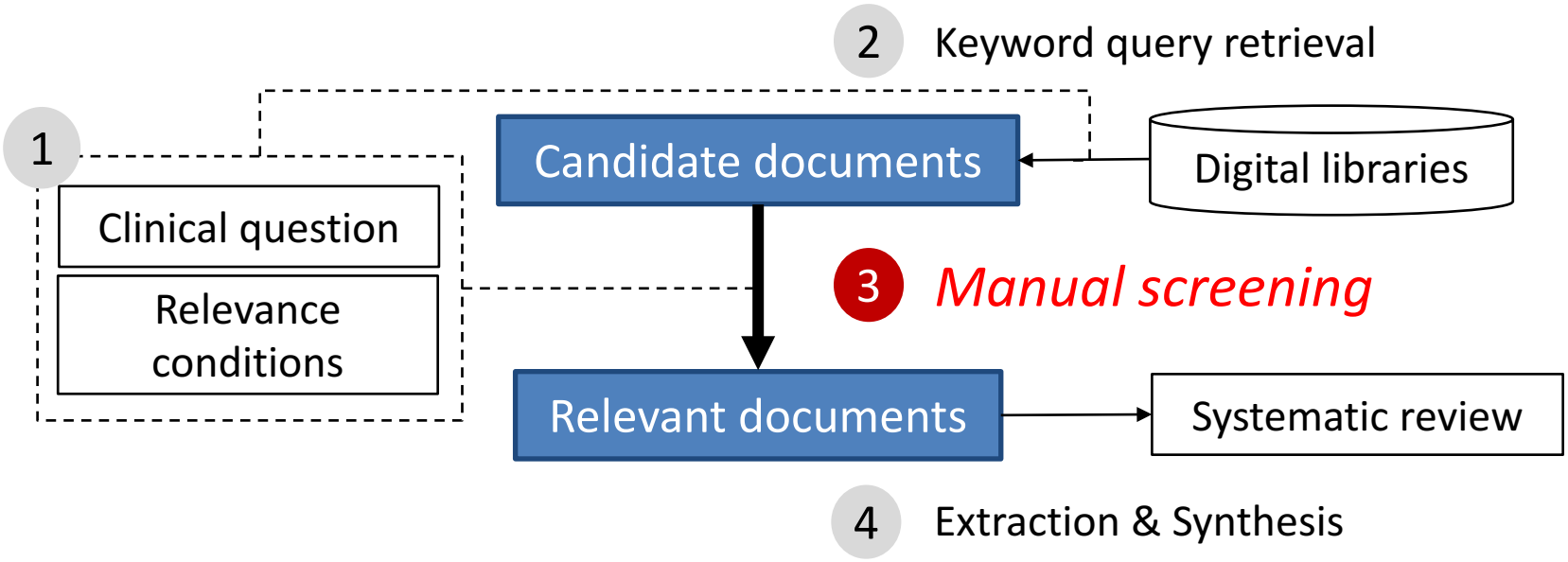


- Extract target data from relevant documents
  - less resource intensive steps

- Example of target data
  - study results
  - experiment methodology
  - subject information

- Analyze and synthesize data to draw an overall conclusion

Our Key Focus



Efficient Screening Process

- Four approaches to improve expensive screening process using text mining

1. Reducing the number of documents to screen
2. Reducing the number of SR experts needed for screening
3. Improving the rate of screening documents
4. Prioritizing the documents to be screened

Screening Prioritization

- Ranked list of candidate documents where relevant documents are at the top
  - SR experts can screen relevant document as early as possible

- Most promising approach to be applied in practice

Alison Ó Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews* 4, 1 (2015), 5.

Seed-driven Document Ranking (SDR)

- New approach for screening prioritization
- Seed document
  - a few relevant documents are known before screening process
  - serve as a query
- *Rank candidate documents where relevant documents are at the top using a seed document*
  - query by document: a long document to short keywords
  - explicit details of document contents
- Understanding characteristics of relevant documents
  - two observations

Our Findings from Analyzing Candidate Documents

- Observation 1.

  For a given SR, its relevant documents share higher pair-wise similarity than that of irrelevant documents.

- Observation 2.

  Relevant documents for a given SR share high commonalities in terms of clinical terms.

  - Unified Medical Language System (UMLS) Metathesaurus
  - extracting clinical terms from the text
    - MetaMap, cTakes, QuickUMLS

Seed-driven Document Ranking (SDR)

- Document representation
  - Observation 2. bag-of-clinical terms (BOC)
  - referring a term to a clinical term

- Weight of a clinical term
  - Observation 1. relevant documents share higher similarities

SDR: Term Weight Method

- Weight $\varphi(t_i, d_s)$ of a clinical term t_i
  - to what extent a term separates similar documents to a seed document, and dissimilar documents

$$\varphi(t_i, d_s) = \ln\left(1 + \frac{\delta(D_{t_i}, d_s)}{\delta(D_{\overline{t_i}}, d_s)}\right)$$

$$\delta(D_*, d_s) = \frac{1}{|D_*|} \sum_{d_j \in D_*} sim(d_j, d_s)$$

- Retrieval model
  - query likelihood model (QLM) with JM smoothing
  - combine the term weight method (φ)

$$score(d, d_s) = \sum_{t_i \in d, d_s} \varphi(t_i, d_s) \cdot c(t_i, d_s) \cdot \log\left(1 + \frac{1-\lambda}{\lambda} \cdot \frac{c(t_i, d)}{L_d \cdot p(t_i|\mathbb{C})}\right)$$

Experiment: Setup

1. Screening prioritization: performance of SDR *(in this presentation)*
   - a single seed document

2. Simulating screening process with SDR
   - multiple labeled relevant documents are available

- Evaluation: average of performances when each relevant document is used as a seed
  - different relevant documents may lead to different performances

Experiment: Data

- CLEF eHealth 2017 (CLEF17) dataset
  - 50 diagnostic test accuracy (DTA) systematic reviews
    - train: 20 SRs
    - test: 30 SRs (competition results)

- Two separated evaluation results
  - test dataset (30 SRs)
  - total dataset (50 SRs)
    - no training in SDR

- Title and abstract of documents
  - clinical term extraction for BOC
  - length of document in BOC: 15% of original document in number of words on average

Experiment: Baselines

- Document representation
  - bag-of-words (**BOW**)
  - bag-of-clinical terms (**BOC**)
- Retrieval model
  - **BM25**
  - query likelihood model (**QLM**)
  - **SDR**

7 models

BM25-BOW QLM-BOW SDR-BOW
BM25-BOC QLM-BOC SDR-BOC
AES

- Average embedding similarity (**AES**)
  - document representation: average of word embeddings
  - ranking score: cosine similarity with a seed document

  - pre-trained word embeddings with PubMed corpus and Wikipedia

Experiment: Evaluation Measures

- Standard IR measures
  - average precision (***avgPr***)
  - precision@*k* (***Pr@k***)
  - recall@*k* (***Re@k***)
    - k = 10, 20, 30

- Task-specific measures
  - normalized *LastRel* by total number of candidate documents (*C*) (***LastRel%***)
    - rank position of last relevant document (*LastRel*)
  - work saved over sampling (***WSS***)

$$WSS = \frac{|C| - LastRel}{|C|}$$

Result: SDR and Baselines

- Result analysis in terms of
  (i) BOC > BOW
  (ii) SDR > AES, BM25, QLM

Dataset	Ranking Model	AvgPr	Pr@10	Pr@20	Pr@30	LastRel%	Re@10	Re@20	Re@30	WSS
30 SRs	CLEF-Query	0.18	-	-	-	46.0	-	-	-	0.54
	BM25-BOW	0.161*	0.176*	0.145*	0.126*	52.9*	0.246*	0.330*	0.385*	0.470*
	QLM-BOW	0.159*	0.165*	0.138*	0.118*	52.0*	0.245*	0.324*	0.376*	0.479*
	SDR-BOW	0.181	0.201*	0.166*	0.139*	46.7	0.257	0.353*	0.401*	0.532*
	BM25-BOC	0.213*	0.233*	0.180*	0.150*	46.5	0.261*	0.345*	0.408*	0.534*
	QLM-BOC	0.214*	0.228*	0.180*	0.150*	43.3*	0.264*	0.361*	0.415*	0.566*
	SDR-BOC	**0.227**	**0.238**	**0.189**	**0.157**	39.8	0.273	**0.367**	**0.436**	0.600
	AES	0.211	0.224	0.175	0.149*	**38.7***	**0.285***	0.364	0.420*	**0.612**
	SDR+AES	***0.264***†‡	***0.276***†‡	***0.213***†‡	***0.177***†	***32.5***†	***0.315***†	***0.413***†	***0.484***†‡	***0.673***†‡
50 SRs	BM25-BOW	0.147*	0.179*	0.146*	0.128*	57.4	0.234	0.305	0.363	0.425*
	QLM-BOW	0.141*	0.168*	0.137*	0.119*	55.7*	0.233*	0.297*	0.343*	0.442*
	SDR-BOW	0.170*	**0.205***	**0.167***	**0.144***	48.5	**0.247**	**0.323**	**0.377**	0.514
	BM25-BOC	0.164*	0.190*	0.151*	0.128*	46.4*	0.230*	0.296*	0.345*	0.535*
	QLM-BOC	0.167*	0.193*	0.156*	0.132*	43.3*	0.233*	0.307*	0.353*	0.567*
	SDR-BOC	**0.178**	0.202	0.164	0.139	**39.8**	0.240	0.312	0.369	**0.601**
	AES	0.147*	0.171*	0.134*	0.115*	50.5	0.238	0.294	0.333*	0.492*
	SDR+AES	***0.202***†‡	***0.226***	***0.179***†‡	***0.152***†‡	***37.7***†‡	***0.265*** †‡	***0.341***†‡	***0.399***†‡	***0.622***†‡

Result: SDR+AES

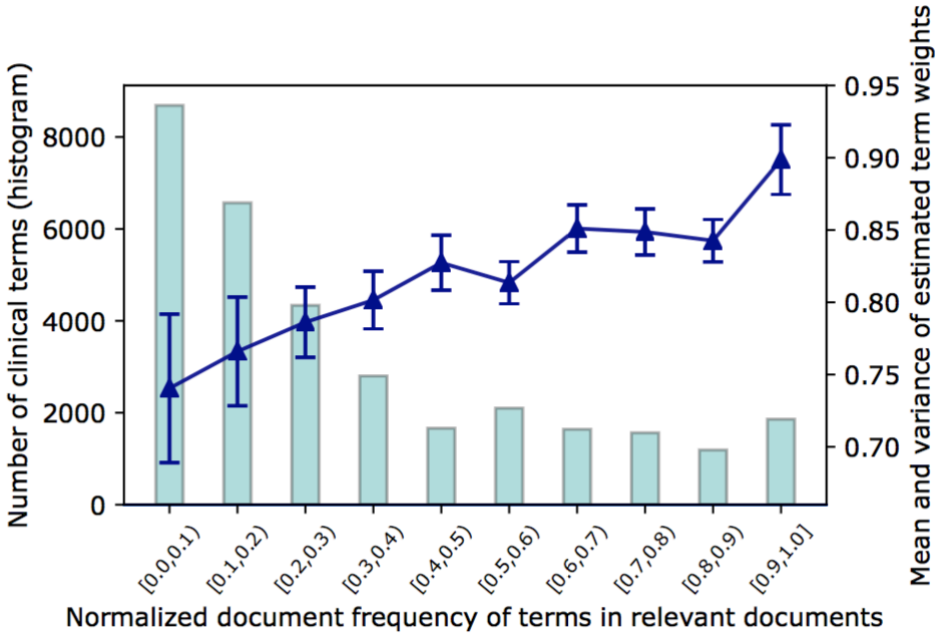
- SDR+AES
  linear combination of ranking scores from SDR-BOC and AES
- SDR-BOC and AES well complement each other

Dataset	Ranking Model	AvgPr	Pr@10	Pr@20	Pr@30	LastRel%	Re@10	Re@20	Re@30	WSS
30 SRs	CLEF-Query	0.18	-	-	-	46.0	-	-	-	0.54
	BM25-BOW	0.161*	0.176*	0.145*	0.126*	52.9*	0.246*	0.330*	0.385*	0.470*
	QLM-BOW	0.159*	0.165*	0.138*	0.118*	52.0*	0.245*	0.324*	0.376*	0.479*
	SDR-BOW	0.181	0.201*	0.166*	0.139*	46.7	0.257	0.353*	0.401*	0.532*
	BM25-BOC	0.213*	0.233*	0.180*	0.150*	46.5	0.261*	0.345*	0.408*	0.534*
	QLM-BOC	0.214*	0.228*	0.180*	0.150*	43.3*	0.264*	0.361*	0.415*	0.566*
	SDR-BOC	**0.227**	**0.238**	**0.189**	**0.157**	39.8	0.273	**0.367**	**0.436**	0.600
	AES	0.211	0.224	0.175	0.149*	**38.7***	**0.285***	0.364	0.420*	**0.612**
	SDR+AES	***0.264***†‡	***0.276***†‡	***0.213***†‡	***0.177***†	***32.5***†	***0.315***†	***0.413***†	***0.484***†‡	***0.673***†‡
50 SRs	BM25-BOW	0.147*	0.179*	0.146*	0.128*	57.4	0.234	0.305	0.363	0.425*
	QLM-BOW	0.141*	0.168*	0.137*	0.119*	55.7*	0.233*	0.297*	0.343*	0.442*
	SDR-BOW	0.170*	**0.205***	**0.167***	**0.144***	48.5	**0.247**	**0.323**	**0.377**	0.514
	BM25-BOC	0.164*	0.190*	0.151*	0.128*	46.4*	0.230*	0.296*	0.345*	0.535*
	QLM-BOC	0.167*	0.193*	0.156*	0.132*	43.3*	0.233*	0.307*	0.353*	0.567*
	SDR-BOC	**0.178**	0.202	0.164	0.139	**39.8**	0.240	0.312	0.369	**0.601**
	AES	0.147*	0.171*	0.134*	0.115*	50.5	0.238	0.294	0.333*	0.492*
	SDR+AES	***0.202***†‡	***0.226***	***0.179***†‡	***0.152***†‡	***37.7***†‡	***0.265*** †‡	***0.341***†‡	***0.399***†‡	***0.622***†‡

Analysis: Term Weight Method

- Calculate normalized DocFreq for clinical terms in relevant documents and bin them into 10 ranges

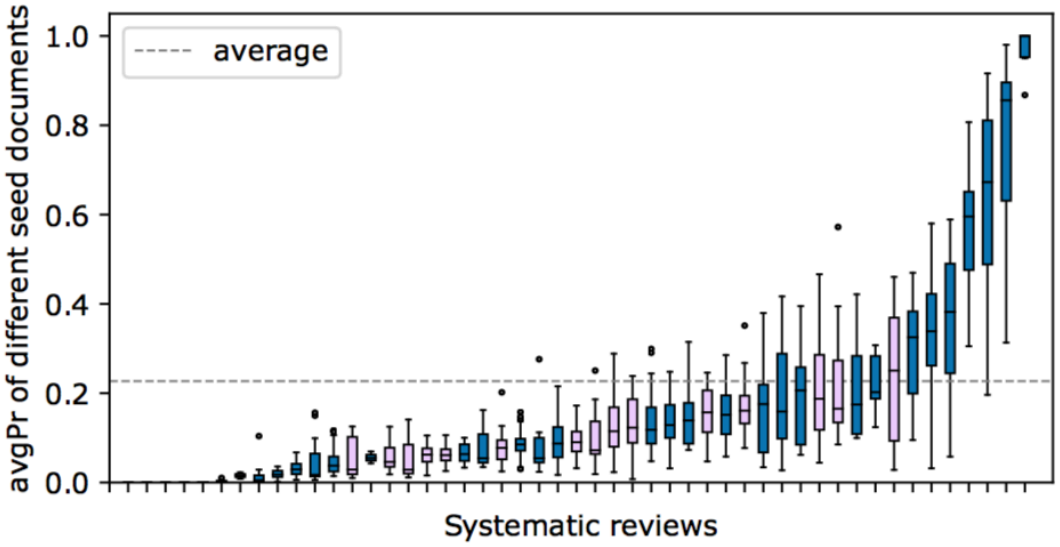
Example. a clinical term appears in all relevant documents: normalized DocFreq = 1.0



- Effective to promote clinical terms which appear in many relevant documents

Analysis: Performance of Individual SRs

- Performance distribution (*avgPr*) of different seed documents within a given SR
  - cause of different performances of SRs: coverage of relevance conditions



- Different difficulties for SRs
- Different performance of seed documents within a SR

Summary

- Seed-driven document ranking (SDR)
  - new approach for screening prioritization
  - domain-specific characteristics
  - seed-driven approach with a weight method
  - extensive analysis of the evaluation results

Thank You!

Thank you for SIGIR Student Travel Grant