American Journal of Information Management
2016; 1(1): 1-9
http://www.sciencepublishinggroup.com/j/infomgmt
doi: 10.11648/j.infomgmt.20160101.11

A Text-Mining Framework for Supporting Systematic Reviews

Dingcheng Li[1, 2], Zhen Wang[1, 3], Liwei Wang[1], Sunghwan Sohn[1], Feichen Shen[1], Mohammad Hassan Murad[3, 4], Hongfang Liu[1]

[1]Department of Health Sciences Research, Mayo Clinic, Rochester, USA

[2]Watson Health Cloud, IBM, Rochester, USA

[3]Robert D. and Patricia E. Kern Centre for the Science of Health Care Delivery, Mayo Clinic, Rochester, USA

[4]Division of Preventive Medicine, Mayo Clinic, Rochester, USA

Email address:

dingcheng.li1@ibm.com (D. Li)

To cite this article:

Dingcheng Li, Zhen Wang, Liwei Wang, Sunghwan Sohn, Feichen Shen, Mohammad Hassan Murad, Hongfang Liu. A Text-Mining Framework for Supporting Systematic Reviews. *American Journal of Information Management.* Vol. 1, No. 1, 2016, pp. 1-9.
doi: 10.11648/j.infomgmt.20160101.11

Received: July 21, 2016; **Accepted**: August 3, 2016; **Published**: August 31, 2016

Abstract: Systematic reviews (SRs) involve the identification, appraisal, and synthesis of all relevant studies for focused questions in a structured reproducible manner. High-quality SRs follow strict procedures and require significant resources and time. We investigated advanced text-mining approaches to reduce the burden associated with abstract screening in SRs and provide high-level information summary. A text-mining SR supporting framework consisting of three self-defined semantics-based ranking metrics was proposed, including keyword relevance, indexed-term relevance and topic relevance. Keyword relevance is based on the user-defined keyword list used in the search strategy. Indexed-term relevance is derived from indexed vocabulary developed by domain experts used for indexing journal articles and books. Topic relevance is defined as the semantic similarity among retrieved abstracts in terms of topics generated by latent Dirichlet allocation, a Bayesian-based model for discovering topics. We tested the proposed framework using three published SRs addressing a variety of topics (Mass Media Interventions, Rectal Cancer and Influenza Vaccine). The results showed that when 91.8%, 85.7%, and 49.3% of the abstract screening labor was saved, the recalls were as high as 100% for the three cases; respectively. Relevant studies identified manually showed strong topic similarity through topic analysis, which supported the inclusion of topic analysis as relevance metric. It was demonstrated that advanced text mining approaches can significantly reduce the abstract screening labor of SRs and provide an informative summary of relevant studies.

Keywords: Systematic Review, Text Mining, Topic Modeling, Keyword Relevance, Indexed-Term Relevance, Topic Relevance, Data Mining

1. Introduction

Evidence-based medicine (EBM) has been shown to play significant roles in informing decision-making regarding the care of individual patients [1]. However, the large number of new publications in health sciences hinder physicians and researchers from keeping up with the latest literature [2]. Therefore, there is a great need for evidence summaries.

Narrative reviews usually involve rapid reviewing so that results can be obtained in a timely manner [3]. For example, at an individual level, busy physicians want to find quick answers from thousands of literatures or at a team level, a group of researchers attempt to acquire current trend of some popular research. In both cases, they may rely on high-reputation journals or highly cited articles to find what they need [4]. However, different from narrative reviews, systematic reviews (SRs) involve a detailed and comprehensive plan and search strategy, with the goal of reducing bias by identifying, appraising, and synthesizing all relevant studies on a particular topic [5]. Therefore, SRs do not rely on journal ranking or abstract-counts to determine whether a study is relevant or not.

High-quality SRs follow strict procedures, and require significant resources and time [6]. At least eight time-consuming steps are needed to conduct a systematic review [7]. Allen and Olkin estimated that a SR with 1000 potential studies retrieved for abstract screening needed 952 working hours to complete [8]. A recent evaluation of 63 SRs conducted by 114 reviewers found that on average a reviewer spent 0.9 minutes, 7 minutes and 53 minutes on abstract screening, full text screening, and data extraction respectively [9, 10]. To keep up to with the latest literature, 7% of SRs needed to be updated at the time of publication, 4% within a year and 11% within 2 years [11]. Therefore, methods that can increase the efficiency of abstract screening without compromising credibility are highly desired.

In this study, we propose a text-mining framework aiming to reduce the burden of screening abstracts in SRs utilizing diverse relevance ranking metrics, including keyword, indexed-term and topic relevance (please see the detailed definition of those relevance metrics in the Methods section). The work to reduce screening burden is fully unsupervised. Meanwhile, all ranking metrics are derived from information retrieval algorithms and offer the flexibility of adding or replacing new ranking metrics. In addition, the framework is highlighted with topic analysis. Specifically, topic analysis, based on Latent Dirichlet Allocation (LDA) [12], is a fully unsupervised model on the basis of word co-occurrences which can group similar documents together. Since its appearance, it has been widely used in natural language processing [13, 14], image processing [15, 16], biomedical informatics [17], and bioinformatics [18] to improve classification [17, 19, 20], summary [21] and other tasks [19, 22]. After conducting the automatic systematic review, we investigated the topic distribution of the abstracts retrieved for each case study in order to find the topic similarities and provide an informative summary.

In the following sections, first related work of this study was introduced, then our approaches described in detail, and finally experiment results were presented using three case studies.

2. Related Works

Attempts to automate abstract screening in SRs started around 2006. O'Mara-Eves et al [23] described the evolution of such approaches and summarized 44 studies that implicitly or explicitly addressed screening workload problems. They concluded that efficiencies and reduction in workload are potentially achievable with text-mining approaches. Across the studies, a saving in workload of 30% -70% was reported as possible using such methods although it may be associated with a loss of 5% of relevant studies (i.e. a 95% recall). Somewhat different from other text-mining applications is that systematic reviewers generally place strong emphasis on high recall (95% to 100%)—that is, a desire to identify all the relevant studies—even if that means a vast number of irrelevant studies need to be considered to find them [23].

Existing automated methods for reducing screening burden in SRs include supervised machine learning and active learning. The task of identifying relevant abstracts can be defined as a binary document classification task where a classifier can be trained to classify abstracts as relevant or irrelevant. Different supervised machine learning algorithms have been explored including the use of naïve Bayes, Adaboost, and SVM by Aphinyanaphongs et al [24, 25], perceptron based voting by Cohen et al [26], factorized version of complement naïve Bayes (FCNB) by Uzuner et al [27], ensembles of SVMs by Wallace et al [28], and evolutionary SVM by Bekhuis and Demner-Fushman [29, 30]. However, supervised machine learning requires annotated training data where informatics researchers rely on existing data gathered in previous SRs. For a given new topic, we may not have previous SRs to serve as training data. In addition, only a small percentage of the abstracts retrieved are relevant which makes the training data very imbalanced. To overcome the above limitations, Wallace et al. [28] and [31] proposed an active online learning approach which starts with a small training set and interactively obtains more training data. To avoid potential overfitting, Jonnalagadda and Petitti [32] incorporated distributional semantics into the active learning process.

In contrast, we consider the task of identifying relevant abstracts as an information retrieval (IR) task with diverse IR relevance ranking metrics considered. Some previous works based on IR have been done [33, 34]. In our proposed SR framework, we also incorporate topic analysis to provide an informative summary as well as to improve relevance ranking. To our knowledge, our work is the first one to integrate topic model into IR approach to reduce screening burden in SRs for new studies. The closest work to ours is Bekhuis et al [35], who built a database of abstracts from 5 systematic reviews and then extracted 5 feature sets from abstracts, including indexing and topic features to train Bayesian classifiers to update relevant articles for previous studies. Two essential differences exist between our proposed approach and theirs. Firstly, they made use of topic probabilities and KL-divergences to generate topic features while we calculate topic relevance with term topic distributions and document topic distributions. Secondly, they focus on finding related new publication leveraging previous studies as training data while we focus on discovering new studies in an unsupervised way.

3. Methods

Figure 1 provides an overview of the proposed framework. The core part of the framework is the three relevance-ranking methods, which are derived from both Query and Topic Analysis. The Query functions as a screening component incorporate diverse IR ranking metrics to rank studies according to their relevance. The Topic Analysis is employed for grouping similar studies together and investigating topic distribution to provide information summary. In the following, details of the framework were provided. Three published Cochrane systematic reviews were used as case studies.

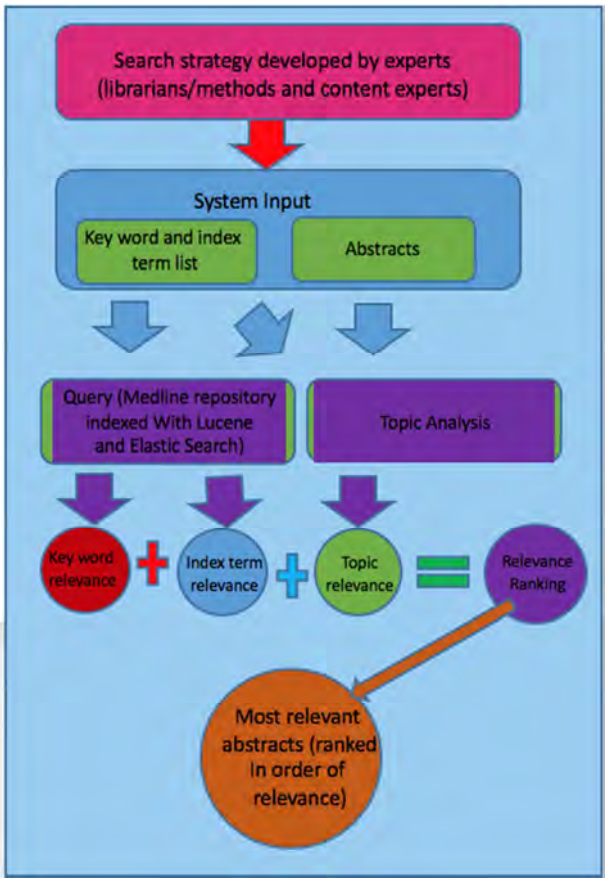

Figure 1. *The text-miningframework for supporting systematic reviews.*

3.1. System Input

The system input includes a list of keywords, the corresponding search strategies adopted from the Cochrane SRs and a list of abstracts retrieved by a librarian for given SR protocols. The keyword list captures important concepts in the SR protocol and is utilized to assess keyword, indexed-term and topic relevance. For both keyword and indexed-term relevance assessment, keywords will be employed as query terms and the search strategies utilized to retrieve abstracts. For topic relevance, weights associated with keywords will be used to compute the relevance score. The abstract list will be used as the collection for identifying relevant studies. In Cochrane SRs, the search strategies are a mix of free text and indexed terms. Since those studies attempt to be comprehensive, diverse databases are involved and search strategies are subtly different for each of them. In this proposed framework, we only utilize a MEDLINE search strategy due to accessibility. In order to see the contributions of each relevance metric, we also separate free text and indexed terms as illustrated below.

3.2. Relevance Ranking

Three semantics-based relevance ranking metrics are named as keyword relevance, indexed-term relevance and topic relevance. Keyword relevance and indexed-term relevance are similar where both measure how relevant an abstract is to the keyword list. Lucene score [36] were adopted to compute the relevance which is based on the term frequency and inverse document frequency (TF-IDF) after screening with a general stop-word list, and is calculated by combining Boolean model and vector space model (VSM) [37]. Specifically, we index the abstract collection using Lucene [36]. A query is then formed by the keyword list. The score returned by Lucene for searching the title and the abstract of the abstract is used to measure keyword relevance. In this combination model, weights obtained by VSM, thresholds are added so that a binary score can be assigned to each weight. Since keywords are generated by users, we may regard keyword relevance as user-defined semantics.

Indexed-term relevance is based on the MeSH terms, which is a comprehensive vocabulary for the purpose of indexing journal abstracts and books in the life sciences provided by the National Library of Medicine (NLM) [38]. Usually, each abstract indexed by PubMed is assigned a group of relevant MeSH terms. Therefore, we suppose that those MeSH terms can reflect the relevance degrees among given studies. The score returned by Lucene for searching the indexed MeSH terms is used to measure the indexed-term relevance. Indexed-term relevance can be different depending on different indexed vocabulary used by different database systems. Indexed terms are usually defined by experts of specific fields. Therefore, this relevance can be also thought as expert-defined semantics.

Topic relevance is derived from topic analysis with LDA, detailed in next section. Each relevance score is normalized across the abstract collection with unit length scaling method to normalize (i.e., $\frac{x}{\sum x}$).

3.3. Topic Analysis with Latent Dirichlet Allocation

In this component, we use the LDA implemented in Mallet Toolkit [39]. All retrieved abstracts and their titles in the case studies are used to construct LDA models. Stop words are removed from the raw documents in a pre-processing step. Then perplexity optimization is used to find the best number of topics where a grid search is made to find the lowest perplexity [12] with the number of topics ranging from 5 to 100. After that certain number of topics are set, and 1000 iterations performed to obtain the topic distributions among given studies. After LDA results are obtained, each topic, represented as a group of words of top probabilities (roughly equivalent to top 10 words) returned by LDA is used to provide high-level information summary. Prominent topics are defined if they cover more than 10% of abstracts.

We assume that studies manually screened tend to have similar topic distributions. Hence, one more relevance metric is defined based on topic distributions and incorporated into the abstract-screening framework. Topic relevance comes from the abstract itself. Specifically, given a query ($\boldsymbol{q}$, the keyword list), the topic relevance score of an abstract (d) is calculated as:

$$P_{lda}(\boldsymbol{q}|d) = \prod_q P_{lda}(q|d, \hat{\theta}, \hat{\phi}) = \prod_q \sum_z P_{\hat{\phi}}(q|z) P_{\hat{\theta}}(z|d),$$

where $\hat{\phi}$ and $\hat{\theta}$ are the posterior estimates of ϕ (the prior of the topic distribution of words) and θ (the prior of the topic distribution of an abstract). In the process, the values of hyper-parameters, α and β need to be determined beforehand. The former controls the abstract distributions while the latter controls the word distributions. The optimal values for α and β can be obtained through grid search as well. Here, we follow the usual heuristic practice [40] by setting α as 50 divided by the number of topics while β as 0.01.

The term $P_{\hat{\phi}}(q|z)$ refers to the probability of the query word given a topic (z) tuned by $\hat{\phi}$ (namely how close query word q to abstract d under topic z). $P_{\hat{\theta}}(z|d)$ refers to the probability of topic z (namely the common hidden semantics of some words or some documents) given abstract d tuned by $\hat{\theta}$. The product of $P_{\hat{\phi}}(q|z)$ and $P_{\hat{\theta}}(z|d)$ refers to how close query word q to abstract d under topic z. The implementation of topic relevance is based on the posterior estimates $\hat{\phi}$ and $\hat{\theta}$, which are outputs from the Mallet.

4. Experiments

4.1. Data Sources

We retrospectively evaluated our framework using three published Cochrane SRs that were chosen to cover different topics (Table 1). The SRs assessed mass media interventions for reducing mental health-related stigma [41], postoperative adjuvant chemotherapy in rectal cancer [42], and the effect of vaccination on preventing influenza in healthy children [43]. The numbers of abstracts retrieved with above-described search strategy from MEDLINE were 3,303, 4,075 and 811 respectively and the numbers after manual screening were 7, 10 and 49 respectively (0.22%, 0.25% and 6% in percentage).

4.2. Evaluation Metrics

We adopted a few metrics that have been utilized previously to measure the screening performance. For a given ranking threshold T, Table 2 provides the definition of each metric.

For a given ranking threshold, the recall change and the reduction in screening burden are the standard metrics used by previous efforts on reducing SR workload [26, 32]. We also pooled the combined effect size of the outcomes using the DerSimonian and Laird random-effect models [44] to show whether meta-analysis estimates derived from results obtained using our framework differ from those in the published Cochrane review (ie, the gold standard list of studies obtained manually). The difference in effect size was tested using the interaction test as described by Altman and Bland [45].

Table 1. *Description of three systematic reviews used as case studies.*

Case	Mass media interventions for reducing mental health-related stigma [41]	Postoperative adjuvant chemotherapy in rectal cancer operated for cure [42]	Vaccines for preventing influenza in healthy children [43]
Objective	To assess the effects of mass media interventions on reducing stigma (discrimination and prejudice) related to mental ill health compared to inactive controls, and to make comparisons of effectiveness based on the nature of the intervention (e.g. number of mass media components), the content of the intervention (e.g. type of primary message), and the type of media (e.g. print, internet).	To quantitatively summarize the available evidence regarding the impact of postoperative adjuvant chemotherapy on the survival of patients with surgically resectable rectal cancer.	To appraise all comparative studies evaluating the effects of influenza vaccines in healthy children, assess vaccine efficacy (prevention of confirmed influenza) and effectiveness (prevention of influenza-like illness (ILI)) and document adverse events associated with influenza vaccines.
Eligibility Criteria	Undergraduate university students from seven upper level psychology courses, two introductory psychology courses, one introductory communications, and two advanced communications	Adults undergoing surgery for rectal cancer who received no adjuvant chemotherapy and those receiving any postoperative chemotherapy regimen.	School children from 2 boarding schools aged 4 to 7 years and 8 to 15 years. There does not appear to be any attrition

The screening performance was assessed for three combinations of relevance metrics by the distribution of relevant studies in five ranking intervals: I. (1-100), II. (100-200), III. (200-300), IV. (300-400) and V. (400, above):

A. keyword relevance

B. linear combination of keyword relevance and indexed-term relevance

C. linear combination of keyword, indexed-term and topic relevance

The interval choice is based on what has been reported in the literature [23] (that is, a saving in workload of between 30% and 70% is expected to be associated with loss of 5% of relevant studies).

Table 2. *Evaluation Metrics Definitions.*

Metrics	**Definition**
Ranking Threshold (T)	Number of abstracts which are used as the threshold.
True positive (TP)	Number of abstract ranking higher than the threshold matching human included studies (this is done by a few professional systematic reviewers)
Recall	The ratio of true positives to the number of relevant studies identified manually

Metrics	Definition
Precision	The ratio of true positives to threshold
Screening saved	The subtraction of total number of abstracts and threshold divided by the total number of abstracts retrieved
Combined effect size	A summary estimate that results from meta-analysis of individual studies included in systematic review.

4.3. Results of Case Studies

Table 3 and Table 4 show the screening performance of our framework and the topic distribution of each case study respectively. Only combination of C was used in Table 3 and Table 4 since it showed the best performance. Although systematic reviewers generally place strong emphasis on high recall, we still report the screening labor for lower recall rates in order to provide a comprehensive view across the three case studies. Figure 2 depicts the proportion of relevant studies for five ranking intervals. In the following, we detailed the results case by case.

Case 1. Mass Media Intervention

The total number of retrieved abstracts is 3,303 and the number of true positives is 7 with the percentage of true positives about 0.2%. When the ranking threshold is 300, we achieved a recall of 100% with 91.8% of the screening labor saved. The ratio of relevant studies in interval I and IIare 0.14 (1 out of 7) and 0.29 (2 out of 7) respectively for A, where only keyword relevance was used. The addition of indexed-term relevance (namely, combination B) brought the inverse proportion for interval I and II (0.29 and 0.14 respectively now). After adding topic relevance (i.e., combination C), there is an increase of 0.43 in ratio for interval I (i.e., increased to 0.72, 5 out of 7).

The number of topics through perplexity optimization was 20. Two prominent topics (4 and 2 abstracts respectively) were found. The top topic words for one (Topic 7) include *brain*, *cortex*, *cognitive* and *temporal* and the other (Topic 17) involves*depression, anxiety, mood* and *suicidal*.

Table 3. *Performance in retrieving relevant studies for three systematic reviews (using all relevant metrics).*

	Case 1 mass media intervention					**Case 2 rectal cancer study**					**Case 3 flu vaccine study**					
Ranking Threshold	3303	400	300	200	100	4075	600	400	300	200	100	811	400	300	200	100
True positives	7	7	7	6	5	10	10	8	7	6	6	49	48	43	33	16
Recall (%)	100	100	100	85.7	71.4	100	100	80	70	60	60	100	98	87.8	67.3	32.3
Precision (%)	0.2	1.8	2.3	3	5	0.25	1.7	2	2.3	3	6	6	12	14.3	16.5	16
Screening saved (%)	0	89	91.8	94.5	97.3	0	85.7	90.2	92.6	95.1	97.5	0	49.3	61.7	73.4	86.3
Combined effect size and 95% confidence interval	0.92	0.92	0.92	0.94	0.96	0.92	0.92	0.93	0.94	0.95	0.96	1.02	1.02	1.01	0.99	0.99
	0.86	0.86	0.86	0.88	0.87	0.86	0.86	0.87	0.88	0.86	0.87	0.94	0.94	0.92	0.92	0.94
	0.99	0.99	0.99	1.02	1.05	0.99	0.99	0.99	1.02	1.01	1.05	1.11	1.11	1.10	1.07	1.05

Table 4. *Topic Distribution of Three Case Studies.*

	Case 1 mass media intervention	**Case 2 rectal cancer study**	**Case 3 flu vaccine study**
Topic No	**Key words (# studies)**	**Key words (# studies)**	**Key words (# studies)**
1	smoking, tobacco, prevalence, cessation	cases, tumor, treated, tissue, bone, years	antibody, vaccine, influenza, hemagglutinin
2	drug, users, abuse, reduction, addiction	survival, patients, rates, surgery, lower	years, age, children, groups, chronic, months (7)
3	studies, trials, interventions, reports	months, medium, relapse, developed	label, respiratory, media, acute (9)
4	adolescents, screening, factors, age (1)	radiotherapy, radiation, rectal, acute	asthma, vaccination, pulmonary, exacerbations (21)
5	participants, weight, increased, trial	tumor, surgical, biopsy, vincristine (4)	virus, antibody, h1n1, h3n2, inhibition, antigen (12)
6	patients, placebo, dose, bseline	mortality, induction, complications, deaths	patients, group, residents, population
7	brain, cortex, cognitive, temporal (4)	cancer, adjuvant, colorectal, adverse	cost, effectiveness, economic, criteria
8	children, behavioral, ratings, families	malignant, progression, brain, surgical	placebo, dose, days, recipients, adults
9	alcohol, survey, questionnaire, questions	prognostic, retrospectively, regression	coverage, increased, persons, data, season
10	women, hiv, sexual, aids, African	chemotherapy, neoadjuvant, pathologic (3)	elderly, high, pandemic, deaths, morbidity
11	internet, web, computer, feedback	carcinoma, pelvic, endometrial, squamous	respiratory, symptoms, fever, illnesses
12	interviews, communication, dementia (1)	lung, patients, cisplation, prospective	reactions, split, immunogenicity, safety
13	psychological, measure, scale, sample	cancer, surgery, therapeutic, oncology	
14	social, autism, fear, examined	breast, tomoxifen, mastectomy, relapse	
15	memory, auditory, attention, motor	resection, liver, metastases, hepatic	
16	mental, public, caregivers, policy	trials, randomized, systematic, advantage (2)	
17	depression, anxiety, mood, suicidal (2)	adjuvant, margins, nodal, invasion	
18	community, prevention, local, based	Complications, performed, laparoscopic	
19	exposure, blood, beta, central, amyloid	dose, fluorouracil, paclitaxel, regimen	
20	disorders, lead, association, diagnostic	trials, adjuvant, randomized, systematic survival	

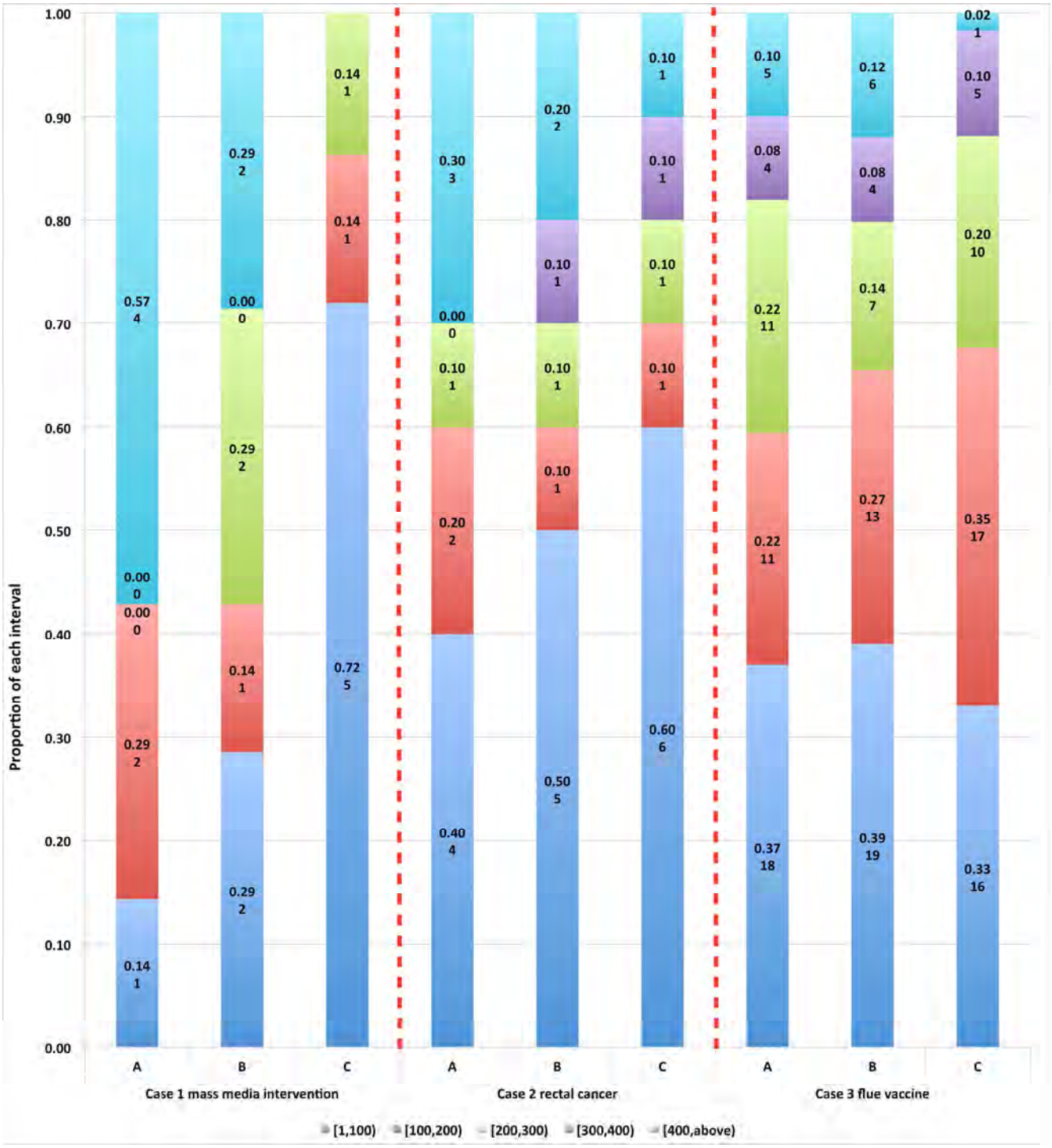

Figure 2. *The proportion of different ranking at each interval for the three case studies. A: keyword relevance. B: linear combination of keyword relevance and indexed-term relevance. C: linear combination of keyword, indexed-term and topic relevance.*

Case 2. Rectal Cancer

The total number of retrieved abstracts is 4,075 and the number of true positives is 10 with the percentage of true positives about 0.25%. When the ranking threshold is 400, we achieved a recall of 80% (8 abstracts, namely) with 90.2% of the screening labor saved. This result did not reach the goal of high recall. Therefore, we also used the threshold 600 for this case, where the recall is 100% and 85.7% is the screening labor saved. The ratio of relevant studies for combination A in interval I is 0.40 (4 abstracts). For combination B, the ratio in interval I increases to 0.50 (i.e. an increase of 0.10, or one more abstract found). The topic relevance (the combination C) brings another 0.10 (another one) increase in interval I (0.60 now or 6 out of 10).

The optimal number of topics through perplexity optimization was 20. The three prominent topics include 4, 3 and 2 abstracts respectively. The top words for the first topic (Topic 5) are clinical related words including *tumor*, *surgical*,

biopsy, *vincristine*, *removal*, *malignant* and *resection*. The second topic (Topic 10) is more therapy related consisting of *chemotherapy, neoadjuvant* and *pathologic*. And the third one (Topic 16) comprises of *trials*, *adjuvant*, *randomized*, *systematic*, *survival* and *regimens*.

Case 3. Influenza Vaccine

For this SR study, the percentage of true positives is about 6% where the number of true positives is 49. A recall of 98% (48 out of 49 abstracts) with 49.3% of the saving in screening labor is achieved when the threshold is 400. For this study, the best ratio achieved in interval I is 0.39 (19 out of 49) by combination B. After adding topic relevance (i.e., combination C), there is a slight decrease of 0.06 in interval I compared to combination C but an increase of 0.08 in interval II. Counting interval I and II, the best results come from combination C (0.59, 0.66 and 0.68 or 29, 32 and 33 for A, B and C respectively)

The optimal number of topics is 12. The 49 relevant studies are mainly distributed among four topics. One topic (Topic 4) includes 21 abstracts where *asthma*, *vaccination*, *pulmonary* and *exacerbations* are dominant words. Another topic (Topic 5) includes *vaccine*, *antibody*, *virus*, *h1n1*, *h3n2* and etc with 12 abstracts in it. The third one (Topic 3) includes 9 abstracts in which *label*, *respiratory*, *media* and *acute* are the top words and the fourth one (Topic 2) includes 7 abstracts with *years*, *age*, *chronic*, *children* and *groups* in it.

5. Discussion

We have described a text-mining framework that reduces the abstract screening burden in SRs while keeping high recall rate and can also provide an informative summary. This framework is partially inspired by our prior work on automated reference assignment [4], which explores methods for assigning reference automatically to expert-written content and also a significant extension of our another work on labor screening reduction [46]. Compared with related work, the proposed framework has multiple advantages. Firstly, it is purely unsupervised. The use of diverse relevance ranking metrics does not require any training data as needed by supervised learning or active learning. Secondly, topic analysis enables the systematic exploration of topics. The topic analysis can be valuable for reviewers to have a better understanding of the relevant studies. Thirdly, our framework has good portability and extensibility. As mentioned in Introduction, we focus on newly conducted SRs while prior works focused on updating existing reviews. However, extension of our framework to update published SRs is possible with minimal effort. We can either run our framework on the newly added studies to test how relevant they are to the previous studies or we can make use of all relevance scores as features to train classifiers. Without doubt, it will be interesting to utilize public resources to make comparisons with other approaches, which will be our future work.

More importantly, the evaluation on three diverse systematic review studies demonstrates robust performance, i.e., adding indexed term relevance and topic relevance boosts the performance comparing to using keyword relevance alone. MeSH terms, as an indexed term system, are derived from experts. It is understandable that MeSH may be a good relevance metric. In Case 1, topic relevance was more helpful than the other two relevance metrics and it brought improvements for both Case 2 and Case 3 as well. Hence, we could say that it is a reasonable relevance considering the unsupervised nature and the modularity of topic modeling. We can flexibly extend topic modeling to incorporate diverse features and to strengthen the model with more representative variables, such as domain knowledge, indexed terms, external resources and so on.

One limitation is that we evaluated our framework retrospectively. To truly assess the contribution of the framework, a prospective indexed study is needed where two groups of systematic reviewers, one with the support of our system and the other following the traditional SR workflow, would conduct demonstrative SRs. The outcome of the two groups can be compared in terms of time spent on abstract screening and the final list of studies selected.

In addition, our current approach for combining relevance metrics is simply an unweighted linear combination. It is noticed that the contribution of relevance metrics for different SR studies is not always consistent. In the future, we plan to give end users options of weighting different relevance ranking metrics.

One other limitation of this study is that only MEDLINE was searched due to accessibility and feasibility issues. It is known that EMBASE [47] and other databases are also important to search in a comprehensive SR. Future work should evaluate text-mining approaches in other databases to enhance portability of proposed frameworks.

A credible SR should summarize evidence from studies selected based on an explicit methodological criteria. Studies should not be selected based on the reputation of journals (impact factor) or authors. Otherwise, the SR would propagate publication bias and not represent the totality of evidence. Therefore, the ranking metrics in our framework (keyword relevance, indexed term relevance or topic relevance) are all purely semantics-based. Potentially, if a rapid (not systematic) review is needed, journal relevance and citation relevance can be used as supplements to our framework.

6. Conclusion and Future Work

It was demonstrated that a text-mining SR supporting framework based on diverse relevance ranking metrics can reduce the labor of SRs to a large degree, while keeping comparably high recall. Meanwhile, we incorporated topic analysis into the framework to provide high level summary of the latest development of intervention trials of given topics. Future work would test such a framework in prospective studies, integrate limited supervision techniques iteratively into SR workflow to further increase recall, and reduce screening burden.

Contributorship Statement

DL led the study design, methodology implementation and drafted the manuscript. DL and FS implemented the data extraction and formation. LW, ZW, MHM, SS and HL gave guidance and consultations on the study designs and on the manuscript editing. HL provided institutional support and manuscript editing. All authors read and approved the final manuscript.

Acknowledgment

The study was supported by the following grants: R01GM102282A1, R01LM11934A1, R01LM11829A1, R01LM11369A1 and 1K99LM012021-01A1.

Special thanks to Yanshan Wang and Yue Yu's help on the final proofreading and editing as well as the strong support from the systematic review group at the Mayo Clinic

References

[1] Serra R, Rizzuto A, Rossi A, Perri P, Barbetta A, Abdalla K, Caroleo S, Longo C, Amantea B, Sammarco G: Skin grafting for the treatment of chronic leg ulcers–a systematic review in evidence‐based medicine. *International wound journal* 2016.

[2] Mulrow CD: Systematic reviews: rationale for systematic reviews. *Bmj* 1994, 309 (6954): 597-599.

[3] Teagarden JR: Meta‐Analysis: Whither Narrative Review? *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 1989, 9 (5): 274-284.

[4] Li D-C, Liu H, Chute CG, Jonnalagadda SR: Towards Assigning References Using Semantic, Journal and Citation Relevance. In: *International Conference on Biomedical Informatics and Biomedicine.* Shanghai, China; 2013.

[5] Uman LS, Chambers CT, McGrath PJ, Kisely S: Psychological interventions for needle-related procedural pain and distress in children and adolescents. *Cochrane Database Syst Rev* 2006, 4.

[6] Higgins JP, Green S: Cochrane handbook for systematic reviews of interventions, vol. 5: Wiley Online Library; 2008.

[7] Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux P, Prasad K, Neumann I, Carrasco-Labra A, Agoritsas T, Hatala R: How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. *JAMA* 2014, 312 (2): 171-179.

[8] Allen IE, Olkin I: Estimating time to conduct a meta-analysis from number of citations retrieved. *Jama* 1999, 282 (7): 634-635.

[9] RxNorm [http://www.nlm.nih.gov/research/umls/rxnorm/]

[10] Wang Z, Noor A, Elraiyah T, Murad M: Dual monitors to increase efficiency of conducting systematic reviews. In: *21st Cochrane Colloquium: Sep 19-23 2013; Quebec, Canada*: Cochrane Collaboration; 2013.

[11] Savoie I, Helmer D, Green CJ, Kazanjian A: Beyond Medline. *International Journal of Technology Assessment in Health Care* 2003, 19 (01): 168-178.

[12] Blei DM, Ng AY, Jordan MI: Latent Dirichlet allocation. *Journal of Machine Learning Research* 2003, 3: 993-1022.

[13] Shu L, Long B, Meng W: A latent topic model for complete entity resolution. In: *Data Engineering, 2009 ICDE'09 IEEE 25th International Conference on: 2009*: IEEE; 2009: 880-891.

[14] Wang X, Mohanty N, McCallum A: Group and topic discovery from relations and text. In*: 2005*: ACM; 2005: 28-35.

[15] Wang C, Blei D, Li F-F: Simultaneous image classification and annotation. In: *Computer Vision and Pattern Recognition, 2009 CVPR 2009 IEEE Conference on: 2009*: IEEE; 2009: 1903-1910.

[16] Wang X, Ma X, Grimson E: Unsupervised activity perception by hierarchical bayesian models. In: *Computer Vision and Pattern Recognition, 2007 CVPR'07 IEEE Conference on: 2007*: IEEE; 2007: 1-8.

[17] Wang H, Ding Y, Tang J, Dong X, He B, Qiu J, Wild DJ: Finding complex biological relationships in recent PubMed articles using Bio-LDA. *PLoS One* 2011, 6 (3): e17243.

[18] Liu B, Liu L, Tsykin A, Goodall GJ, Green JE, Zhu M, Kim CH, Li J: Identifying functional miRNA–mRNA regulatory modules with correspondence latent dirichlet allocation. *Bioinformatics* 2010, 26 (24): 3105-3111.

[19] Li D, N Xia, S Sohn, KB Cohen, CG Chute, H Liu: Incorporating Topic Modeling Features For Clinic Concept Assertion Classification. In: *The 5th International Symposium on Languages in Biology and Medicine (LBM 2013) 12th and 13th Decemeber, 2013 2013; Tokyo, Japan*; 2013.

[20] Ogilvie MM, Tearne CF: Spontaneous abortion after hand-foot-and-mouth disease caused by Coxsackie virus A16. *British medical journal* 1980, 281 (6254): 1527.

[21] Liu H, Wang T, Wei Y, Zhao G, Su J, Wu Q, Qiao H, Zhang Y: Detection of type 2 diabetes related modules and genes based on epigenetic networks. *BMC Syst Biol* 2014, 8 Suppl 1: S5.

[22] Li D, T Thermeau, CG Chute, H Liu: Discovering Associations Among Diagnosis Groups Using Topic Modeling. In: *AMIA Summits on Translational Science Proceedings: 2013; San Francisco*; 2013.

[23] Alison O, Thomas J, McNaught J, Miwa M, Ananiadou S: Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews* 2015, 4 (1): 5.

[24] Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF: Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association* 2005, 12 (2): 207-216.

[25] Cohen AM: Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@ 95 measure. *Journal of the American Medical Informatics Association* 2011, 18 (1): 104-104.

[26] Cohen AM, Hersh WR, Peterson K, Yen P-Y: Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 2006, 13 (2): 206-219.

[27] Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR: Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association* 2012: amiajnl-2011-000784.

[28] Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH: Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics* 2010, 11 (1): 55.

[29] Bekhuis T, Demner-Fushman D: Towards automating the initial screening phase of a systematic review. *Stud Health Technol Inform* 2010, 160 (Pt 1): 146-150.

[30] Bekhuis T, Demner-Fushman D: Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers. *Artificial intelligence in medicine* 2012, 55 (3): 197-207.

[31] Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S: Reducing systematic review workload through certainty-based screening. *Journal of biomedical informatics* 2014, 51: 242-253.

[32] Jonnalagadda S, Petitti D: A new iterative method to reduce workload in systematic review process. *International Journal of Computational Biology and Drug Design* 2013, 6 (1): 5-17.

[33] Lee E, Dobbins M, DeCorby K, McRae L, Tirilis D, Husson H: An optimal search filter for retrieving systematic reviews and meta-analyses. *BMC medical research methodology* 2012, 12 (1): 1.

[34] Shirahatti A: Text Retrieval for Systematic Reviews.

[35] Bekhuis T, Tseytlin E, Mitchell KJ, Demner-Fushman D: Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. *Plo Sone* 2014, 9 (1): e86277.

[36] Hatcher E, Gospodnetic O, McCandless M: Lucene in action. In.: Manning Publications; 2004.

[37] Pérez-Iglesias J, Pérez-Agüera JR, Fresno V, Feinstein YZ: Integrating the probabilistic models BM25/BM25F into Lucene. *arXiv preprint arXiv:09115046* 2009.

[38] Lowe HJ, Barnett GO: Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Jama* 1994, 271 (14): 1103-1108.

[39] McCallum AK: {MALLET: A Machine Learning for Language Toolkit}. 2002.

[40] Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P: The author-topic model for authors and documents. In: *Proceedings of the 20th conference on Uncertainty in artificial intelligence: 2004*: AUAI Press; 2004: 487-494.

[41] Clement S, Lassman F, Barley E, Evans‐Lacko S, Williams P, Yamaguchi S, Slade M, Rüsch N, Thornicroft G: Mass media interventions for reducing mental health‐related stigma. *The Cochrane Library* 2013.

[42] Petersen SH, Harling H, Kirkeby LT, Wille‐Jørgensen P, Mocellin S: Postoperative adjuvant chemotherapy in rectal cancer operated for cure. *The Cochrane Library* 2012.

[43] Jefferson T, Rivetti A, Harnden A, Di Pietrantonj C, Demicheli V: Vaccines for preventing influenza in healthy children. *The Cochrane Library* 2008.

[44] DerSimonian R, Laird N: Meta-analysis in clinical trials. *Control Clin Trials* 1986, 7 (3): 177-188.

[45] Altman DG, Bland JM: Interaction revisited: the difference between two estimates. *Bmj* 2003, 326 (7382):219.

[46] Li D, Z Wang, F Shen, MH Murad, H Liu: Reducing the Screening Burden of Systematic Review with a Multiple-level Relevance Ranking System. In: *American Medical Informatics Association: 2014; Washington, DC*; 2014.

[47] Woods D, Trewheellar K: Medline and Embase complement each other in literature searches. *Bmj* 1998, 316 (7138): 1166.