RESEARCH ARTICLE

Recommending research articles to consumers of online vaccination information

Eliza Harrison[1], Paige Martin[1], Didi Surian[1], and Adam G. Dunn[1,2]

[1]Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
[2]Discipline of Biomedical Informatics and Digital Health, School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, Australia

Keywords: information retrieval, news media, research communications, vaccination

ABSTRACT

Online health communications often provide biased interpretations of evidence and have unreliable links to the source research. We tested the feasibility of a tool for matching web pages to their source evidence. From 207,538 eligible vaccination-related PubMed articles, we evaluated several approaches using 3,573 unique links to web pages from Altmetric. We evaluated methods for ranking the source articles for vaccine-related research described on web pages, comparing simple baseline feature representation and dimensionality reduction approaches to those augmented with canonical correlation analysis (CCA). Performance measures included the median rank of the correct source article; the percentage of web pages for which the source article was correctly ranked first (recall@1); and the percentage ranked within the top 50 candidate articles (recall@50). While augmenting baseline methods using CCA generally improved results, no CCA-based approach outperformed a baseline method, which ranked the correct source article first for over one quarter of web pages and in the top 50 for more than half. Tools to help people identify evidence-based sources for the content they access on vaccination-related web pages are potentially feasible and may support the prevention of bias and misrepresentation of research in news and social media.

1. BACKGROUND

The communication of health and medical research information online provides a critical resource for the public. More than three quarters of the UK public report an interest in biomedical research, with 42% having actively sought out content relating to medical or health research in 2015 (Huskinson, Gilby, et al., 2016). Nearly all searches for health information take place online via search engines (Castell et al., 2014; Fox & Duggan, 2013; Fox & Rainie, 2002; Huskinson et al., 2016). Internet searches are a common way for people to engage with health research and the communication of health research on news websites and other forums, and have the potential to alter health beliefs and decisions (Weaver, Thompson, et al., 2009).

The communication of health research in news and social media is associated with several challenges. Studies with fewer participants and of lower methodological rigor are more common in news media (Haneef, Ravaud, et al., 2017; Selvaraj, Borkar, & Prasad, 2014), and research from authors with conflicts of interest tends to receive more attention in news and social media (Grundy, Dunn, et al., 2018). As many as half of all news reports manipulate or sensationalize study results to emphasize the benefits of experimental treatments (Yavchitz, Boutron, et al., 2012).

an open access journal

Citation: Harrison, E., Martin, P., Surian, D., & Dunn, A. G. (2020). Recommending research articles to consumers of online vaccination information. *Quantitative Science Studies*, *1*(2), 810–823. https://doi.org/10.1162/qss_a_00030

DOI: https://doi.org/10.1162/qss_a_00030

Received: 01 April 2019
Accepted: 11 February 2020

Handling Editor: Vincent Larivière

Corresponding Author: Adam G. Dunn
adam.dunn@sydney.edu.au

Copyright: © 2020 Eliza Harrison, Paige Martin, Didi Surian, and Adam G. Dunn. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

The MIT Press

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

Despite issues with the reliability of health information online, most people trust what they encounter (Fox & Rainie, 2000, 2002), and are inconsistent in their efforts to validate health information using appropriate sources (Eysenbach, 2002; Fox & Rainie, 2002), likely because they find it difficult to do so. Where attempts to assess the credibility of health information are made, the visibility and accessibility of sources such as scientific research articles are an important criterion by which users assess the quality of online health communications (Eysenbach, 2002; Fox & Rainie, 2002). Individuals are also subject to order-effect biases that impact their perception of the evidence presented by online communications of health research (Lau & Coiera, 2007), and tend to believe information that aligns with their current knowledge of a health topic (Fox & Rainie, 2002).

The representation of medical research in the public domain is particularly important in relation to vaccination, where vocal critics actively seek to erode trust in the safety and effectiveness of vaccines and immunization programs. In 2019, the World Health Organization listed vaccine hesitancy—the reluctance or refusal to vaccinate—as one of the 10 most significant threats to global health (World Health Organization, 2019). There is a clear risk that the misrepresentation of scientific evidence and amplification of misinformation by social media may be major contributing factors to further outbreaks of these diseases in future (Larson, 2018).

The rise of vaccine hesitancy as a global public health issue is in part driven by the increased pervasiveness of antivaccination sentiment in search engine results (Kata, 2012) and the mainstream news media (Larson, Cooper, et al., 2011), as well as the growth of social media as a platform for the provision of a diverse range of information sources to the public (Steffens, Dunn, & Leask, 2017). Discussion of the safety and efficacy of vaccines is a common theme in news reports and low-quality information is common (Cooper Robbins, Pang, & Leask, 2012). On web pages specifically advocating against vaccination, the majority cite safety risks, including illness, damage, or death (Bean, 2011; Kata, 2010).

To effectively identify biases and misrepresentation in online articles that communicate the outcomes of health research, we need to be able to quickly identify the original source literature for said research. While existing services such as Altmetric (https://www.altmetric.com/) can be used to identify links to scientific source material using Digital Object Identifiers (DOIs), Uniform Resource Locators (URLs), or other identifiers such as PubMed IDs (PMIDs), in most cases these identifiers must be embedded in hyperlinks to enable their tracking. Other media services that offer more complete tracking of media mentions of research tend to be for-profit subscription services that support organizations wanting to keep track of their research outputs. These services are source centric—they start with a research article and track the media that reference it—and may not easily support use cases where a member of the public is interested in accessing the source research that underpins the information on web pages communicating health-related research to the public.

Our aim was to evaluate methods for automatically identifying source literature by recommending articles for web pages communicating vaccination research to the public. To do this, we made use of a large set of reported links between vaccination-related web pages and the scientific literature they reference tracked by Altmetric.

2. METHODS

2.1. Study Data

The study data comprised a set of research articles from PubMed linked to a set of web pages via Altmetric. To construct the corpus of research articles from PubMed, we retrieved all

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

articles from PubMed by searching for "*vaccine,*" automatically expanded to include searches for the plural form, and "*vaccine*" as a Medical Subject Heading (MeSH) term. Title and abstract text for each article were extracted using the National Center for Biotechnology Information (NCBI) E-Utilities Application Programming Interface (API) (https://www.ncbi.nlm.nih.gov/books/NBK25501/). Any PubMed articles that did not include at least 100 words after concatenating title and abstract were excluded from the analysis, and the remaining articles formed the PubMed corpus (Figure 1). The search was conducted in July 2018.

We then used the Altmetric API to identify the set of research communications that linked to one or more of the articles in the PubMed corpus. We defined research communications to include news articles, blogs, and non-social-media posts that discuss the outcomes of vaccine-related research. After crawling each URL to access the web articles, contiguous blocks of text

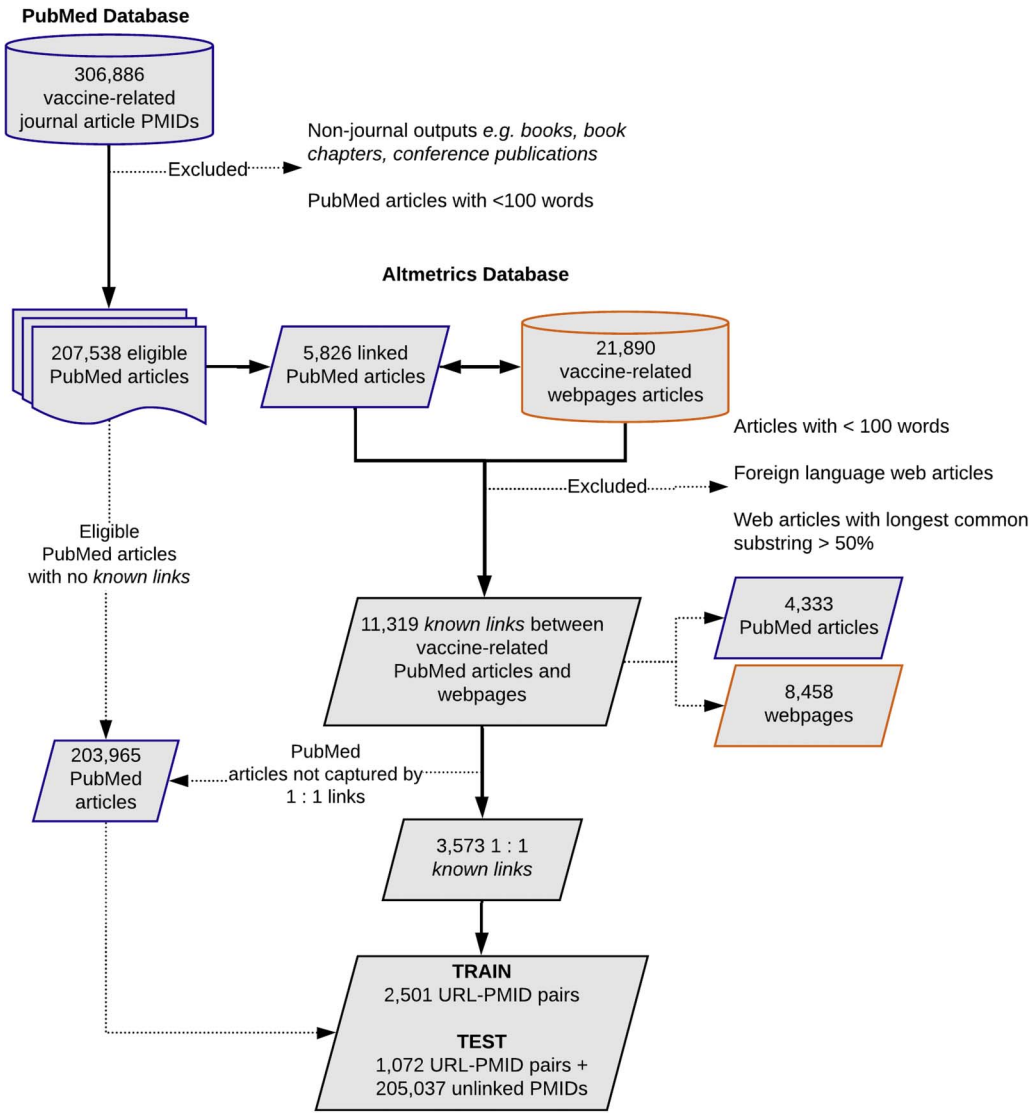

Figure 1. The process for the collection and processing of the study data sets. Of the eligible journal articles retrieved from the PubMed database we identified 11,319 distinct *known links* (URL-PMID pairs) corresponding to research outputs and vaccine-related web pages with links tracked by Altmetric, 3,573 of which were used to train and test the proposed approaches.

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

from the web pages were concatenated to form the basis of the data used in the following analyses. Text from the set of web pages was accessed in July 2018. Web pages were excluded if they did not include at least 100 words of text, as were any identified as non-English using the Google Code language-detection library (https://code.google.com/p/language-detection/). We also excluded web pages with substantial amounts of exact duplicate text and which referenced a single source research article. This was common where articles were published by multiple local news platforms owned by a single entity, often with only minor changes in title, content, or formatting. To remove these duplicates, we identified web pages for which the longest common substring between any two records linked to a PMID was greater than 50% of the total length of the longest web page. We then randomly selected web pages such that no PMID was mapped to any number of similar web pages. Note that after selecting unique examples of linked web pages and research articles, no two web pages had a longest common substring overlap of more than 10% of the total length.

The resulting data set included 207,538 research articles, of which 4,333 had known links to one or more of 8,458 distinct web pages (Figure 1). There were 1,934 articles that were referenced on two or more web pages, with one article referenced by 98 distinct web pages. Conversely, there were 1,418 web pages that referenced two or more research articles, one of which had known links to 68 of the articles in the PubMed corpus. To generate a final set of reported links for which no web page was linked to more than one PubMed article in the final corpus and vice versa, we first selected any article and web page pairs for which the corresponding PMID and URL were both present only once in the data set (1:1 links). For each of the remaining research articles, we instead selected the linked web page with the greatest number of words that was not yet present in final corpus. This resulted in a final set of 3,573 PMID-URL pairs of 1:1 linked articles and web pages, which we refer to as the *known links* set.

2.2. Feature Extraction and Dimensionality Reduction

To generate a term-based vector representation of each of the linked articles and web pages, we preprocessed each document by removing punctuation and words consisting entirely of numeric characters. We then used the remaining words to construct a vocabulary of terms common to both corpora (terms that existed in at least one research article and at least one web page).

Each article or web page was then represented as a vector of numeric values based on one of three standard vector representations: *binary*, *term frequency* (TF), and *term frequency-inverse document frequency* (TF-IDF). Binary vectors were generated by recording the presence (value = 1) or absence (value = 0) of vocabulary terms in each document. The TF vector representation was defined as a count of the number of times each word appeared in the document. The TF-IDF score is given by the log-transformed TF value multiplied by the inverse of the log-transformed proportion of documents in which the feature was present. In contrast to term frequency, TF-IDF weights vary depending on how common the term is across the entire corpus, based on the assumption that words appearing more often in fewer documents (like the name of a specific vaccine or the outcomes measured in a research study) are likely to be more informative, while those that appear often across many documents (like "and," "the," or "vaccination") are less informative (Spärck Jones, 1972; Ramos, 2003; Robertson, 2004).

In information retrieval methods, sparse representations of documents may be less useful for measuring document similarity or finding documents relevant to a search. This is expected in particular for short documents. To address issues of sparsity, dimensionality reduction methods transform the representation of a document into fewer dimensions.

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

We evaluated the use of two approaches. The first was a simple feature reduction method that uses threshold parameters. Features were removed by applying the maximum document frequency limit of 0.85 to the combined corpora vocabulary. As a result, those terms common to more than 85% of articles and web pages in the corpus were excluded from the term-based vector representation.

For the second dimensionality reduction approach we used *truncated singular value decomposition* (T-SVD). T-SVD works in a similar way to singular value decomposition (SVD) by decomposing a matrix into a product of matrices that contain singular vectors and singular values. The singular values can be used to understand the amount of variance in the data captured by the singular vectors. T-SVD allows more efficient computation than SVD since T-SVD approximates the decomposition by only considering a select few components, specified as an argument to the algorithm (Halko, Martinsson, & Tropp, 2011).

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

2.3. Ranking Methods

We used cosine similarity as a standard measure of similarity between web pages and PubMed articles. For each web page, we calculated the cosine similarity to all 205,037 articles in the test portion of the final document corpus to produce a ranked list.

We expected that there would be consistent differences between the language style used in article titles and abstracts, compared to that used in online research communications. For example, we expected that communications would replace technical jargon with simpler synonyms. *Canonical correlation analysis* (CCA) (Hotelling, 1936) is an algorithm designed to identify linear combinations of maximally correlated variables between complex, multivariate data sets. CCA captures and maps the correlations between two sets of variables into a single space, and thus the comparison for ranking can be made using a standard similarity measure. CCA is used to analyze a joint dimensionality reduction across different spaces (e.g., text and images, text and text, etc.) (Menon, Surian, & Chawla, 2015; Rasiwasia, Costa Pereira, et al., 2010). As a result, the CCA approach could be used to learn the alignment between the terms used in the articles and the terms used to describe the same concepts in research communications presented online. To test the CCA approach, we added it as an extra process in the pipeline, using training data to construct a transform (a matrix that may modify the number of features), and then apply that transform to the testing data before calculating the distance (Figure 2).

2.4. Experiments and Outcome Measures

While standard document similarity methods typically do not need to be constructed on one set of data and tested on another, the CCA approach learns an alignment between articles and web pages based on a set of training data, and its ability to generalize to unseen data is best tested on a separate data set. To examine the effect of adding CCA to the pipeline, we constructed *training* and *testing* sets by randomly assigning each PMID-URL pair. The resulting training data set comprised 70% (or 2,501) of the known links, with the remaining 30% of PMID-URL pairs allocated to the testing set. To replicate the work of searching a large corpus or database for relevant scientific publications, we also added the 203,965 eligible PubMed articles not already captured in either the training or testing data sets, resulting in a testing set of 1,072 linked articles and web pages plus the set of 203,965 articles with no linked web pages.

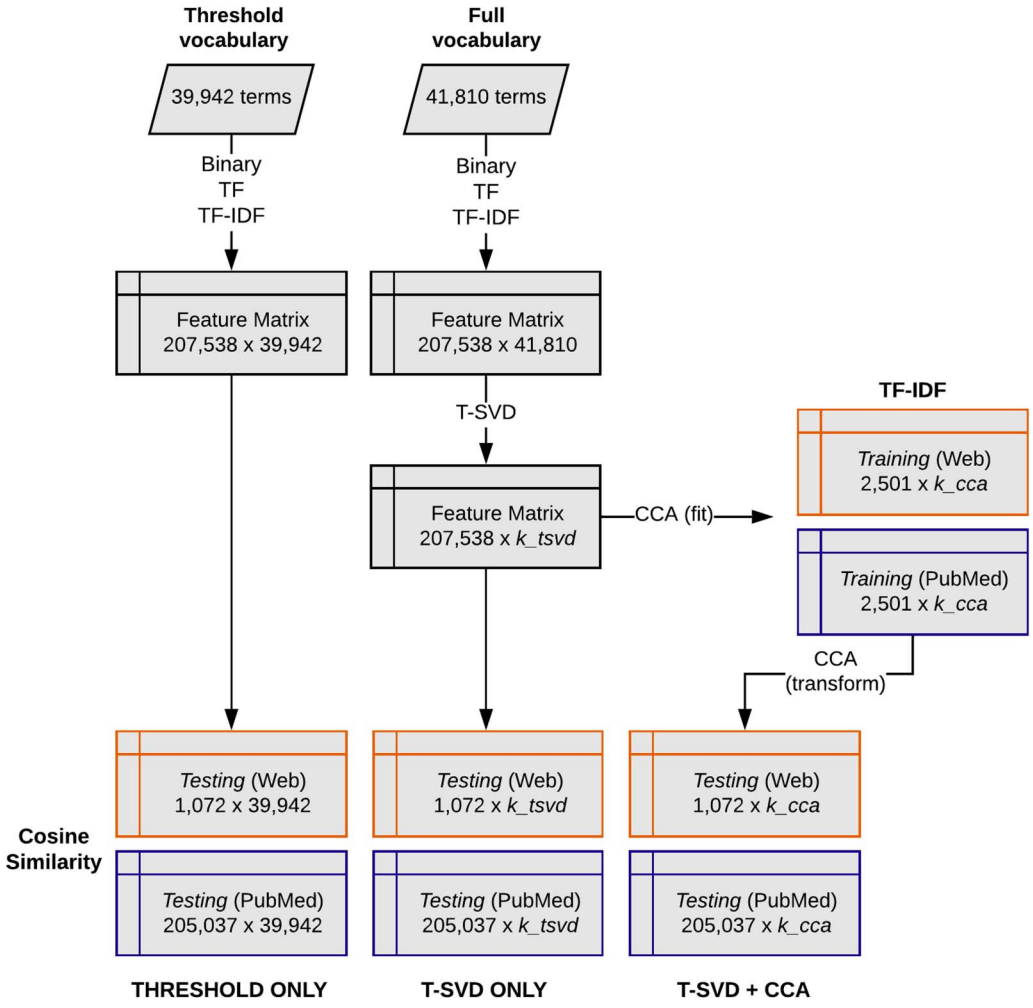

Figure 2. We compared three methods of representing terms in using a vocabulary reduced using maximum document frequency parameters (*threshold only*) or that reduced using T-SVD (*T-SVD only*), the performance of each was measured by ranking the cosine similarity values between each web page and article in the testing set. Also tested was the effect of transforming the best performing feature representation using both T-SVD and CCA on document similarity rankings (*T-SVD + CCA*).

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

The set of experiments were split into two phases. In the first phase, we examined how differences in the vector space representations might affect the performance of the ranking methods, comparing the binary, TF, and TF-IDF representations in combination with either threshold or T-SVD feature reduction. In the second, we tested the effect of transforming the best performing feature representation using CCA.

The success of each of these systems in correctly linking research articles to the web pages that reference them is indicated by the final rank of the correct PubMed article for each of the 1,072 web pages tested. Based on the similarity between each web page and source article we calculated the number of PubMed articles that a user would be required to read to locate the known links for at least half of all web pages, equivalent to the median rank of the correct source article. As a second metric we determined the number of web pages for which the correct PubMed article was ranked first out of all possible 205,037 articles in the testing set, or the

proportion of known links correctly identified by each system (i.e., recall@1). We also calculated the proportion of links ranked within the top 50 PubMed articles in the testing set as an indicator of the capacity of each system to return the correct PubMed article within the first page of query results (i.e., recall@50). Finally, we plotted recall@k for all values between 1 and the total number of PubMed articles to visualize the proportion of known links that can be identified after having read the top *k* ranked source articles.

All methods and experiments were developed using Python 3.6, the code for which is available on GitHub (https://github.com/evidence-surveillance/web2pubmed). Data for the final set of pairs used in the evaluation are also available in the same repository.

3. RESULTS

Among the 207,538 articles that were returned by the search and met the inclusion criteria for the analysis, 4,333 had one or more links to web pages recorded by Altmetric and were also eligible for inclusion in study analyses. The most popular article was used as source information on 98 web pages, while 22% (2,535 of 11,319 known links) were used as source information on one web page (Figure 2). To construct a representative data set in which no article or web page was represented more than once, we selected a final set of 3,573 PMID-URL pairs.

Within this final set of 3,573 articles and web pages with known PMID-URL links and 203,965 additional articles with no known links, we identified 41,810 terms used at least once in both the set of web pages and the set of articles. Where we applied threshold parameters (limiting the vocabulary to exclude terms used in at least 85% of corpus documents), this vocabulary was reduced to 39,942 terms, representing the greatest number of features used in the following analyses. For experiments instead using the T-SVD method of feature reduction, the number of terms retained in the data set varied between 100 and 1,600.

Of the methods of representing the text of articles and web pages, we observed that TF-IDF consistently produced the highest performance (Table 1). Regardless of the feature reduction approach used, experiments using the TF-IDF representation of document text outperformed the binary and TF representations.

Of the two feature reduction methods, the threshold approach outperformed the T-SVD approach for all outcome measures (Table 1). However, because the performance improved roughly linearly as the number of T-SVD components was increased, the results suggest that the number of features used may be a more important factor than the choice of feature reduction method. Overall, the highest performance was achieved using TF-IDF to represent the text as term features and document frequency thresholds to reduce the number of features. In the testing data set, this method ranked the correct source article first for more than one in four web pages and placed the correct source article in the top 50 ranked candidate articles for more than half of the web pages.

The addition of CCA was expected to improve the performance of the method by finding an alignment between the terms used in the web pages and articles rather than exact matches between terms. We found that adding CCA to the process improved the performance for experiments where the number of T-SVD components was relatively low (Table 2). However, as we increased the number of T-SVD components above 400, the improvements gained from adding CCA started to diminish, indicating that the maximum gain in performance from adding CCA was achieved for the experiment that used 400 T-SVD components transformed into 200 feature dimensions by the trained CCA model, where for 38.0% of the web pages, the

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

Table 1. Performance of document similarity methods in a set of 205,037 PubMed articles

Feature representation and reduction methods	Median rank (IQR)	Recall@1	Recall@50
Threshold parameters			
Binary	238.5 (1–9154)	0.25	0.42
TF	427.5 (5–10075.25)	0.19	0.37
TF-IDF	**41 (1–799.25)**	**0.26**	**0.52**
T-SVD (100 components)			
Binary	8858 (1198–34252.25)	0.05	0.10
TF	38491.5 (4968.75–104229.25)	0.05	0.08
TF-IDF*	2768 (203.5–24884.5)	0.07	0.17
T-SVD (200 components)			
Binary	5522.5 (495–27377.5)	0.07	0.14
TF	36429 (3924.75–99717)	0.05	0.09
TF-IDF*	1513 (84.75–15572.25)	0.10	0.22
T-SVD (400 components)			
Binary	3211.5 (188–21040.25)	0.09	0.18
TF	31220 (2967.25–96203.5)	0.07	0.10
TF-IDF*	720 (36–9674.25)	0.13	0.28
T-SVD (800 components)			
Binary	1606 (41.75–15311.75)	0.13	0.26
TF	29421 (2245.25–92871.5)	0.07	0.12
TF-IDF*	385.5 (13–6211.25)	0.15	0.34
T-SVD (1600 components)			
Binary	824.5 (9–12704.5)	0.17	0.33
TF	29519.5 (1597.5–93890)	0.08	0.13
TF-IDF*	219 (6–4145.75)	0.17	0.37

* Experiments for which results have also been included in Table 2.

IQR: interquartile range; TF: term frequency; TF-IDF: term frequency-inverse document frequency; T-SVD: truncated singular value decomposition.

correct source article was placed within the top 50 ranked candidates (Figure 3). As the number of feature dimensions used was increased further, the approach then failed because the CCA failed to converge because of the sparsity of the feature space (Figure 4). Overall, the results show that we were able to identify a maximum performance within the parameter space for which the CCA approach could be used, but that none outperformed the simpler approach that used thresholds rather than T-SVD and did not use CCA (Figure 5).

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

Table 2. CCA-based alignment methods for identifying the evidence source for vaccination web pages among a set of 205,037 candidate articles in the testing data set.

Method (CCA dimensions)	Median rank (IQR)	Recall@1	Recall@50
100 T-SVD components			
No CCA*	2768 (203.5–24884.5)	0.07	0.17
50	318.0 (23.0–3381.0)	0.10	0.32
100	475.0 (20.0–4635.5)	0.10	0.32
200 T-SVD components			
No CCA*	1513 (84.75–15572.25)	0.10	0.22
50	322.5 (20.0–2940.0)	0.09	0.31
100	200.0 (10.0–1982.75)	0.13	0.36
200	253.5 (11.0–4198.0)	0.14	0.36
400 T-SVD components			
No CCA*	720 (36–9674.25)	0.13	0.28
50	575.0 (60.0–5055.5)	0.05	0.23
100	268.5 (15.0–2696.5)	0.10	0.32
200	**185.5 (7.0–2506.75)**	0.14	**0.38**
400	270.0 (11.0–5581.0)	0.14	0.35
800 T-SVD components			
No CCA*	385.5 (13–6211.25)	0.15	0.33
50	3806.5 (279.75–28002.75)	0.02	0.12
100	1100.0 (29–15787.0)	0.03	0.21
200	409.0 (27.0–10816.0)	0.08	0.30
400	291.5 (15.0–9859.0)	0.11	0.34
800	1437.0 (34.0–34434.75)	0.08	0.27
1600 T-SVD components			
No CCA*	219 (6–4145.75)	**0.17**	0.37
50	58164.5 (19678.0–117859.25)	0.00	0.01
100	47806.0 (14104.5–110966.5)	0.00	0.02
200	37414.5 (7236.75–92341.25)	0.00	0.04
400	30554.5 (3454.0–91052.25)	0.00	0.06
800	NA†	NA	NA
1600	NA†	NA	NA

* Experiments for which results also appear Table 1.

† Experiments in which the CCA did not converge.

IQR: interquartile range; CCA: canonical correlation analysis; t-SVD: truncated singular value decomposition.

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

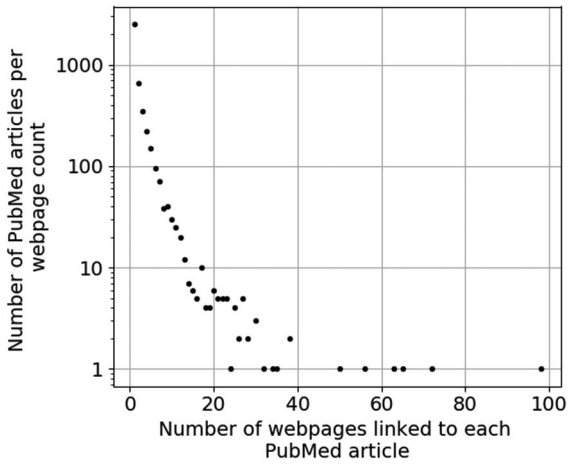

Figure 3. The distribution of the number of distinct web pages (URLs) linked to each vaccine-related article retrieved from PubMed (PMIDs).

4. DISCUSSION

In this study we evaluated methods that could be used as part of tools to support the identification of missing links between online research communications and the source literature they use. We used vaccination research as an example application domain where there are common problems with bias and misrepresentation in subsequent news and media coverage. We started with the assumptions that many web pages are not reliably connected to the research on which they are based, and that readers may not have the time or expertise to construct a search query to identify relevant articles in bibliographic databases. We tested methods that seek to circumvent the need for expert construction of search queries and instead automatically recommend articles that are likely to be relevant. While the use of a CCA-based approach did not outperform our baseline methods, the results suggest that such tools are likely feasible.

4.1. Methods for Automatic Recommendations from Text

We tested two standard information retrieval methods and found that the simpler approach using a TF-IDF representation and a maximum document frequency limit outperformed a more

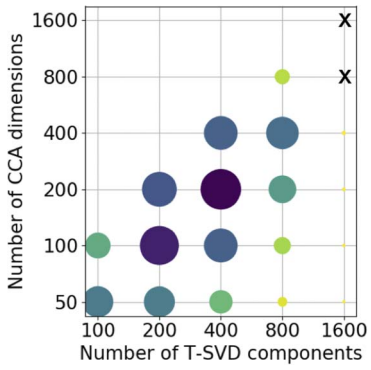

Figure 4. A visual comparison of the difference in inverse median rank (circle area) for each of the experiments varying the number of T-SVD components and the number of CCA dimensions. T-SVD experiments where the CCA did not converge are marked with a cross.

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

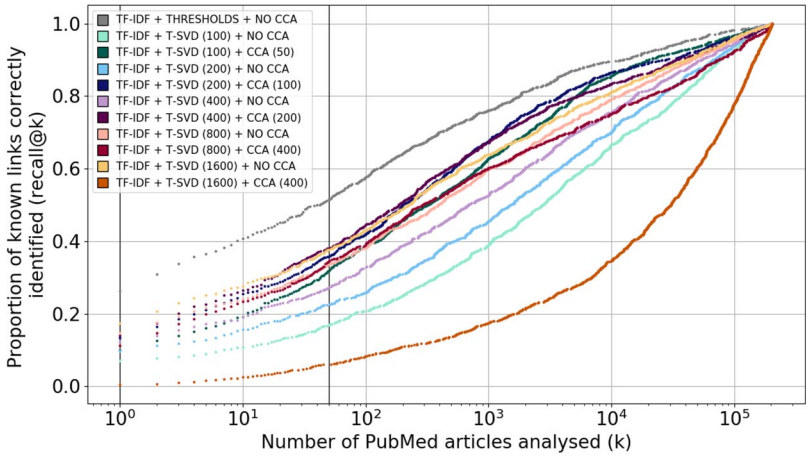

Figure 5. The recall@k for experiments comparing the addition of CCA (darker colors) to no CCA (lighter colors), varying the number of T-SVD components and compared to the best performing baseline approach.

sophisticated approach of transforming the feature space using CCA. While we know of no previous studies that have developed tools for the same purpose, the structure of the problem is common. The combination of TF-IDF and cosine distance has previously been used to identify missing links between trial registrations on ClinicalTrials.gov and articles in PubMed reporting trial results (Dunn, Coiera, & Bourgeois, 2018). Similarly, the use of TF-IDF and T-SVD has been shown to facilitate the detection of similarities between patent documents and scientific publications (Magerman, van Looy, & Song, 2010). These results were consistent with ours—increasing the number of SVD components improved the accuracy, but the best performance was achieved without the use of SVD.

There are a range of other more complex approaches that could be applied to a problem of this structure: the identification of missing links between two distinct sets of documents that may be matched using similarity of content and a relatively sparse bipartite graph connecting the two sets of documents. These might include alternative feature representations, such as pretrained language models, word embedding, or both (Beam, Kompa, et al., 2020; Howard & Ruder, 2018; Mikolov, Sutskever, et al., 2013; Peters, Neumann, et al., 2018), as well as other algorithms for recommendation or ranking related to collaborative filtering (Huang, Li, & Chen, 2005; Koren, Bell, & Volinsky, 2009), and learning-to-rank methods (Ibrahim & Landa-Silva, 2017; Liu, 2009).

An expert might take an alternative approach to manually identifying source articles for online research communications, making use of specific information, including the names of authors, institutions, or journals. Rule-based approaches that make use of this information may yield improvements. Other similar approaches might make use of the date of publication extracted from web pages and articles in bibliographic databases, under the assumption that online communications of research tend to be reported soon after the research is published.

4.2. Implications and Future Applications

The results indicate that it is likely feasible to build a tool that could be used to help find missing links between health research communications and source literature for the purpose of checking the veracity of the communications and identifying biases. One way to operationalize this type of tool would be to develop browser plugins that automatically augment web

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

pages with a list of recommended relevant peer-reviewed research. Hyperlinks might be added to the terms or phrases that most contribute to the recommendation, based on the weights of the terms that contribute to the similarity.

A further application relates to the automatic detection of distortion or bias in research communications. Checklist tools such as QIMR (Zeraatkar, Obeda, et al., 2017) or DISCERN (Charnock & Shepperd, 2004; Charnock, Shepperd, et al., 1999) are designed to be used to manually evaluate the credibility of health information and health research communications, but little work has been done to use these checklists as the basis for automatically estimating the credibility of web pages (Shah, Surian, et al., 2019). We know of no studies that have attempted to automatically compare the text of research communications with the abstract or full text of research articles to detect specific differences that might be indicative of misrepresentation of distortion of research conclusions. For example, tools able to identify scenarios where studies of association are written as causation in communications would be of clear benefit, particularly when discussing vaccination (Kata, 2012; Moran, Lucas, et al., 2016).

Tools extending the work we present here could also be used to help educate nonexperts on when it is appropriate to search for source articles when reading research communications online, and to train them on how to construct useful search queries. First, the distances to the top-ranked articles might be suggestive of whether the text on a web page is based on any form of peer-reviewed research. This could be used to indicate a common practice in antivaccine blogs, where writers provide circular links within a network of other blogs that are all equally disconnected from clinical evidence. Second, the tool could be used to show users a search query that is automatically generated from the text of research communications for use with bibliographic databases like PubMed, educating users on how to search bibliographic databases for clinical evidence.

4.3. Limitations

This study had several limitations. First, while the use of Altmetric helped us to quickly construct a large data set of reported links, the data set might be a biased sample of web pages that communicate about vaccination research. Communications that include hyperlinks to journal web pages or PubMed or that link to articles using their DOIs may be of higher quality or may be targeted at specialized audiences. Other online research communications not using hyperlinks may be different from those tracked by Altmetric. Testing the approaches on a more general set of examples before deployment would be necessary. Second, there are a wide range of alternative approaches to feature representation and recommender systems. While we discuss the potential advantages of some of these approaches above, we are at present only able to speculate on which of them are likely to perform best as part of a tool or service aimed at improving the detection of distortion in research communications online. Finally, while vaccination is an important application domain, we did not test what might happen if we had selected a much broader sample of web pages and articles, or if we had constructed models specifically designed to find missing links for individual fields or topics of research. It is possible that more general or more specific data sets may influence the performance of the methods we tested.

5. CONCLUSION

The results indicate the feasibility of tools designed to support the identification of missing links between health research communications and the scientific literature on which they are based. Such tools have the potential to help people better discern the veracity and quality

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

of what they read online. While standard feature representation and document similarity methods were moderately successful in this task, further investigation is warranted.

AUTHOR CONTRIBUTIONS

Eliza Harrison: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—original draft, Writing—review & editing. Paige Martin: Data curation, Writing—review & editing. Didi Surian: Conceptualization, methodology, Writing—review & editing. Adam Dunn: Conceptualization, Methodology, Supervision, Project Administration, Writing—review & editing.

COMPETING INTERESTS

There are no competing interests to declare.

FUNDING INFORMATION

Macquarie University Postgraduate Scholarship (Eliza Harrison).

DATA AVAILABILITY

Metadata, including the set of 3,573 linked PubMed identifiers and web page URLs, are provided in a GitHub repository (https://github.com/evidence-surveillance/web2pubmed) alongside the code used to perform experiments.

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022

REFERENCES

Beam, A. L., Kompa, B., Fried, I., Palmer, N. P., Shi, X., Cai, T., & Kohane, I. S. (2020). Clinical concept embeddings learned from massive sources of medical data. *Pacific Symposium on Biocomputing*, *25*, 295–306.

Bean, S. J. (2011). Emerging and continuing trends in vaccine opposition website content. *Vaccine*, *29*(10), 1874–1880. https://doi.org/10.1016/j.vaccine.2011.01.003

Castell, S., Charlton, A., Clemence, M., Pettigrew, N., Pope S., Quigley, A., Shah, J. N., & Silman, T. (2014). *Public attitudes to science 2014: Main report. URN BIS/14/P111, Ipsos MORI*.

Charnock, D., & Shepperd, S. (2004). Learning to DISCERN online: Applying an appraisal tool to health websites in a workshop setting. *Health Education Research*, *19*(4), 440–446. https://doi.org/10.1093/her/cyg046

Charnock, D., Shepperd, S., Needham, G., & Gann, R. (1999). DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *Journal of Epidemiology and Community Health*, *53*(2), 105–111. https://doi.org/10.1136/jech.53.2.105

Cooper Robbins, S. C., Pang, C., & Leask, J. (2012). Australian Newspaper Coverage of Human Papillomavirus Vaccination, October 2006–December 2009. *Journal of Health Communication*, *17*(2), 149–159. https://doi.org/10.1080/10810730.2011.585700

Dunn, A. G., Coiera, E., & Bourgeois, F. T. (2018). Unreported links between trial registrations and published articles were identified using document similarity measures in a cross-sectional analysis of ClinicalTrials.gov. *Journal of Clinical Epidemiology*, *95* (Mar), 94–101. https://doi.org/10.1016/j.jclinepi.2017.12.007

Eysenbach, G. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ*, *324*(7337), 573–577. https://doi.org/10.1136/bmj.324.7337.573

Fox, S., & Duggan, M. (2013). *Health Online 2013. Pew Internet & American Life Project*. https://www.pewresearch.org/internet/2013/01/15/health-online-2013/

Fox, S., & Rainie, L. (2000). The online health care revolution. *Pew Internet & American Life Project: Online Life Report*. https://www.pewresearch.org/internet/2000/11/26/the-online-health-care-revolution/

Fox, S., & Rainie, L. (2002). Vital decisions: A Pew Internet Health Report. *Pew Internet & American Life Project*. https://www.pewresearch.org/internet/2002/05/22/vital-decisions-a-pew-internet-health-report/

Grundy, Q., Dunn, A. G., Bourgeois, F. T., Coiera, E., & Bero, L. (2018). Prevalence of disclosed conflicts of interest in biomedical research and associations with journal impact factors and altmetric scores. *JAMA*, *319*(4), 408. https://doi.org/10.1001/jama.2017.20738

Halko N., Martinsson P. G., & Tropp J. A. (2011). Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, *53*(2) (May), 217–288. https://doi.org/10.1137/090771806

Haneef, R., Ravaud, P., Baron, G., Ghosn, L., & Boutron, I. (2017). Factors associated with online media attention to research: A cohort study of articles evaluating cancer treatments. *Research Integrity and Peer Review*, *2*(9), 1–8. https://doi.org/10.1186/s41073-017-0033-z

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*(3/4), 321. https://doi.org/10.2307/2333955

Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia. https://doi.org/10.18653/v1/P18-1031

Huang, Z., Li, X., & Chen, H. (2005). Link prediction approach to collaborative filtering. In *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries—JCDL '05*, pp. 141–142. New York, NY: ACM Press. https://doi.org/10.1145/1065385.1065415

Huskinson, T., Gilby, N., Evans, H., Stevens, J., & Tipping, S. (2016). Wellcome Trust Monitor Report Wave 3 Tracking public views on science and biomedical research. *Wellcome Trust Monitor: Wave 3*. https://wellcome.ac.uk/sites/default/files/monitor-wave3-full-wellcome-apr16.pdf

Ibrahim, O. A. S., & Landa-Silva, D. (2017). ES-Rank: Evolution strategy learning to rank approach. In *Proceedings of the Symposium on Applied Computing—SAC '17*, pp. 944–950. New York, NY: ACM Press. https://doi.org/10.1145/3019612.3019696

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine*, *28*(7), 1709–1716. https://doi.org/10.1016/J.VACCINE.2009.12.022

Kata, A. (2012). Anti-vaccine activists, Web 2.0, and the postmodern paradigm—An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*, *30*(25), 3778–3789. https://doi.org/10.1016/J.VACCINE.2011.11.112

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, *42*(8), 30–37. https://doi.org/10.1109/MC.2009.263

Larson, H. J. (2018). The biggest pandemic risk? Viral misinformation. *Nature*, *562*(7727), 309–309. https://doi.org/10.1038/d41586-018-07034-4

Larson, H. J., Cooper, L. Z., Eskola, J., Katz, S. L., & Ratzan, S. (2011). Addressing the vaccine confidence gap. *The Lancet*, *378*, 526–535. https://doi.org/10.1016/S0140

Lau, A. Y. S., & Coiera, E. W. (2007). Do people experience cognitive biases while searching for information? *Journal of the American Medical Informatics Association*, *14*(5), 599–608. https://doi.org/10.1197/jamia.M2411

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, *3*(3), 225–331. https://doi.org/10.1561/1500000016

Magerman, T., van Looy, B., & Song, X. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, *82*(2), 289–306. https://doi.org/10.1007/s11192-009-0046-6

Menon, A. K., Surian, D., & Chawla, S. (2015). Cross-modal retrieval: A pairwise classification approach. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pp. 199–207. Philadelphia, PA: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974010.23

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *In NIPS '13 Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2*, pp. 3111–3119.

Moran, M. B., Lucas, M., Everhart, K., Morgan, A., & Prickett, E. (2016). What makes anti-vaccine websites persuasive? A content analysis of techniques used by anti-vaccine websites to engender anti-vaccine sentiment. *Journal of Communication in Healthcare*, *9*(3), 151–163. https://doi.org/10.1080/17538068.2016.1235531

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of NAACL-HLT 2018*, pp. 2227–2237.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R. G., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia—MM '10*, pp. 251–260. New York, NY: ACM Press. https://doi.org/10.1145/1873951.1873987

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, Piscataway, NJ. https://www.cs.rutgers.edu/~mlittman/courses/ml03/iCML03/papers/ramos.pdf

Robertson, S. (2004) Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520. https://doi.org/10.1108/00220410410560582

Selvaraj, S., Borkar, D. S., & Prasad, V. (2014). Media coverage of medical journals: Do the best articles make the news? *PLoS ONE*, *9*(1), e85355. https://doi.org/10.1371/journal.pone.0085355

Shah, Z., Surian, D., Mandl, K. D., & Dunn, A. G. (2019). Automatically applying a credibility appraisal tool to track vaccination-related communications shared on social media. *Journal of Medical Internet Research*, *21*(11), e14007. https://doi.org/10.2196/14007

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Steffens, M., Dunn, A. G., and Leask, J. (2017). Meeting the challenges of reporting on public health in the new media landscape. *Australian Journalism Review*, *39*(2), 119–132.

Weaver, J. B., Thompson, N. J., Weaver, S. S., & Hopkins, G. L. (2009). Healthcare non-adherence decisions and internet health information. *Computers in Human Behavior*, *25*(6), 1373–1380. https://doi.org/10.1016/J.CHB.2009.05.011

World Health Organization (WHO). (2019). Ten threats to global health in 2019. Retrieved March 1, 2019, from https://www.who.int/emergencies/ten-threats-to-global-health-in-2019

Yavchitz, A., Boutron, I., Bafeta, A., Marroun, I., Charles, P., Mantz, J., & Ravaud, P. (2012). Misrepresentation of randomized controlled trials in press releases and news coverage: A cohort study. *PLOS Medicine*, *9*(9), e1001308. https://doi.org/10.1371/journal.pmed.1001308

Zeraatkar, D., Obeda, M., Ginsberg, J. S., & Hirsh, J. (2017). The development and validation of an instrument to measure the quality of health research reports in the lay media. *BMC Public Health*, *17*(1), 343. https://doi.org/10.1186/s12889-017-4259-y

Downloaded from http://direct.mit.edu/qss/article-pdf/1/2/810/1885763/qss_a_00030.pdf by guest on 19 August 2022