Text mining for search term development in systematic reviewing: a discussion of some methods and challenges

Claire Stansfield. Alison O'Mara-Eves, James Thomas

Abstract

Using text mining to aid the development of database search strings for topics described by diverse terminology has potential benefits for systematic reviews; however, methods and tools for accomplishing this are poorly covered in the research methods literature. We briefly review the literature on applications of text mining for search term development for systematic reviewing. We found that the tools can be used in five overarching ways: improving the precision of searches; identifying search terms to improve search sensitivity; aiding the translation of search strategies across databases; searching and screening within an integrated system; and developing objectively-derived search strategies. Using a case study and selected examples, we then reflect on the utility of certain technologies (TF-IDF and Termine, term frequency, and clustering) in improving the precision and sensitivity of searches. Challenges in using these tools are discussed. The utility of these tools is influenced by the different capabilities of the tools, the way the tools are used, and the text that is analysed. Increased awareness of how the tools perform facilitates the further development of methods for their use in systematic reviews.

Keywords: text mining, information retrieval, systematic search, clustering

Background and aims

Database searching is a core requirement when undertaking many systematic reviews, and the choice of search terms used is key in identifying relevant literature in a systematic way. Identifying search terms to locate an unknown body of literature is challenging, particularly for literature that uses diverse terminology or is not consistently indexed. For example, in a literature review about services and systems to promote the self-care of minor ailments, a range of conceptual perspectives and vocabulary describes 'self-care', including 'self-help', 'seeking information', 'treat at home' (Richardson et al., in press). Developing (combinations of) search terms for this type of review is often an iterative process, which can be aided by analysing patterns in samples of text in order to assess which words or phrases can capture relevant studies, and to find ways to minimise the number of irrelevant studies retrieved. The process is imprecise and database searches are normally supplemented by other search methods including the checking of reference lists and citations, and contacting key informants. However, the evolution of sets of search terms can involve a variety of techniques, including knowledge of the literature, published pre-existing searches in related areas, topic expertise, database thesauri, iterative searching, browsing citations within databases, and – the focus of this paper – text mining.

'Text mining' in this paper describes a variety of processes that enable discovery of words and patterns in collections of text. Advantages of using text mining tools for a literature search may include: supporting the scanning of a large corpus of preliminary results for identification of keywords and subject terms with the potential to improve search strategies; improving time efficiency; and, in some cases, providing a reproducible, objective method (as opposed to human-developed search strategies that rely on experience or knowledge of the users) (Paynter et al., 2016). There are indications from the literature on search filter development that user-derived 'intuitive' search terms are not always the most suitable terms for using in a search strategy (Petrova et al., 2012; White et al., 2001) and analysing text or relevant citations might be useful in countering this problem. For example, text mining helped to identify relevant search terms for a systematic review on the broad topic of community engagement, beyond the list of terms that the authors developed themselves (O'Mara-Eves et al., 2014). Two analyses within the separate disciplines of health (Hausner et al., 2015) and software engineering (Zhang et al., 2011) found that text mining can be used to obtain studies that were not obtained from researcher-derived search terms in the original search strategies. However, Hausner et al. (2015) found that both user-derived approaches and an objective approach that relied on text mining for informing the search terms could each miss some relevant references in reviews on certain non-drug intervention topics. Thomas et al. (2011, p.4) observe that, through text mining, "the range of search terms can be expanded in a way that better describes the literature in the review", However, they also point out "its limitation is a function of its strength: it expands the review in favour of the literature that uses the same language as the documents that have already been found".

Paynter et al. (2016) undertook an overview of text mining tools and techniques in systematic reviews and identified 111 tools, of which 52 support searching. They concluded that "Although it seems promising, text mining has not become a standard tool for creating systematic review search strategies" (p.13), and noted one possible limitation was that many tools have been developed based upon output from PubMed or Medline. These databases are typically used in development because they are large, well-structured, open datasets. However, systematic reviewers often need to work across many databases, and databases differ in how they structure citations and controlled vocabularies. This may reduce the generalisability of tools that have been developed based on limited datasets, as they may not transfer well to other databases (and domains).

It is important to be aware that the application of text mining for search strategy development is distinct from the related area of search filter development that requires considerable investment in terms of developing gold standard sets of literature upon which to build and test filters. In contrast with search filters, search strategies for specific systematic reviews are often developed for specific reviews and need to be developed relatively quickly. However, there are lessons from filter development using text mining that can be applied to search strategy development. For some specific topic areas, developing filters using word frequency analysis is challenging, and sometimes impossible, as shown by attempts to develop a filter on road safety interventions (Wendt et al., 2001) and health-related social values (Petrova et al., 2012). Some search filters for topics that are described by diverse terminology have been developed combining both text mining and expert knowledge or manual processes, for example, alcohol-impaired driving (Goss et al., 2007), prognosis of work disability (Kok et al., 2015) and overviews of systematic reviews (Lunny et al., 2016). In addition, terminology also needs to be considered in context. For example Kok et al. (2015) and Petrova et al. (2012) observed their topic search filters behaved differently across different health conditions.

Given the challenges in creating search filters from representative samples of literature on a topic, text mining is considered here as an aid rather than a complete solution for informing search strategies for topics that encompass a range of conceptual perspectives or are described by varied vocabularies. This complementary approach also mitigates potential bias from the sample of literature used for text mining, which may only help identify more of the same literature. There seems to be a paucity of published literature on text-mining procedures for identifying free-text and controlled terms for specific databases using generic tools, though EUnetHTA (2015) and Gourlay (2010) provide some guidance on obtaining term frequencies. Controlled vocabularies can also be analysed using database specific tools, particularly for Medline and PubMed, and these are listed elsewhere (Paynter et al., 2016, HLWIKI Canada contributors, 2016). However, there is little guidance on utilising the variety of text mining tools available to complement other methods to identify search terms for undertaking systematic reviews.

The aims of this paper are to: 1) give an overview of the main applications of text mining for search term development; 2) reflect on the usefulness of some technologies through a case study and further examples; and 3) discuss the challenges in using these tools. We hope this will promote further debate and dissemination of techniques and methods.

An overview of applications of text mining for search term development

The purpose of this section is to provide an overview of the main applications of text mining for search term development. To do this, examples of the application of text mining for search term development in reviews were identified from the following sources: items screened for a systematic review on text mining for screening (O'Mara-Eves et al., 2015); focused iterative searches of Google and Google Scholar; citation searches of literature found; browsing the repository SRtoolbox.com; and discussion groups, such as the Cochrane Information Retrieval Methods Group.

The types of applications identified for text mining for search term development are shown in Box 1. These show five groups: increasing the sensitivity (or recall) of a search; increasing the precision of a search; aiding translation across databases; searching and screening within an integrated system; and using text mining as the predominant method for 'objective searches'. Objective searches are outside of our focus here, and are described by Hausner et al. (2012; 2015; 2016).

Box 1 Applications of text mining

Increasing sensitivity: Identifying more words, word forms or phrases (O'Mara-Eves et al., 2014; Zhang et al., 2011)
Increasing precision: Identifying combinations of words (Thompson et al., 2014) or phrases; identifying words from clustering to 'safely' exclude terms at low risk of missing studies (Stansfield et al., 2013)
Aiding translation across databases: Identifying free-text terms from records that would not be captured by the controlled terms (Damarell et al., 2013);
Search and screening within an integrated repository system (Mergel et al., 2015)
Developing objective search strategies, where all the search terms are derived from a suitable sample (Hausner et al., 2012; 2015; 2016; Simon et al., 2010)

Text mining can be used to increase the *sensitivity* of a search by identifying more words, word forms or phrases, to broaden the range of studies that contain relevant records (Damarell et al., 2013; O'Mara-Eves et al., 2014; Zhang et al., 2011). This might be targeted on certain elements of the search, for example, Damarell et al. (2013) identified potential search terms in the titles and abstracts from records only retrieved by a database's controlled vocabulary and not by known free-text terms.

The *precision* of a search can be improved by identifying phrases or combinations of words rather than a single word on its own, such as Thompson et al. (2014), or by identifying themes of unwanted items through automated clustering (Stansfield et al., 2013).

For *aiding translation* across databases, Damarell et al. (2013) used text mining to capture items from PubMed not indexed with controlled Medical Subject Headings; however, this could have wider applications in assisting development of search strategies across other databases. The reverse of this approach is also used where citations identified from searches of free-text fields in a database are analysed for suitable controlled vocabularies.

Mergel et al. (2015) describe SLRqub, as a proof of concept, as a tool to enable search query-building of the software engineering research repository, IEEExplore Library. The tool uses the results from a search and manual assessment by the user of relevant and non-relevant studies, in order to suggest search terms, and facilitate further searching and screening within the repository. Such an approach could be possible for reviews in different disciplines, once the included studies have been determined, though it is likely to be more resource-intensive, particularly where many databases have been searched, and it may be difficult to apply if the search terms are made up of multiple components.

Although presented here as distinct applications, these applications could be utilised in combination and iteratively during search term development, or perhaps to analyse sub-components of a search. They could also potentially be used on a set of screened records, either as quality assurance (O'Mara-Eves et al., 2013), or as part of an integrated searching and screening system (Mergel et al., 2015).

Text mining tools used in the following case study and examples

In the case study and examples below, text mining was applied in two of the five applications from our framework above: improving the precision of searches, and identifying search terms to improve sensitivity by determining both useful and undesirable search terms and phrases to help refine the search strategy. Text mining was used as a part of designing the search strategies, but significant human input was also involved in designing the search, choosing search terms, running test searches and browsing results.

Generic text mining tools, which are not reliant on datasets for specific databases, were used. These were readily available to explore their utility, and do not require specialist computer science support to use. They were identified in various ways: from EPPI-Reviewer, professional networks and browsing the literature. We consider these tools to represent some fairly standard analytical options that are currently available.

We start by describing briefly various categories of text mining tools and situate the tools mentioned in this paper within these categories. The text mining tools explored here centre around three distinct types of technology: term frequency; automatic term recognition; and automatic clustering.

Term frequency involves obtaining frequencies of word occurrence and co-occurrence. *BibExcel* (Perrson n.d.) can generate a word list showing how many citations contain specific words (for example, the separate words 'disability', 'disabled'), the stem of a word (for example, 'disab') and co-occurring words. Some reference management software can be used to generate lists of controlled vocabulary rapidly through its subject bibliography function (Hayman & Shaheem, 2014); in the case study, we used *Endnote*. Concordance tools, such as *AntConc* (Anthony, 2014) can reveal collocates (words within a certain distance of other words) and N-grams (sequences of n words) within large volumes of text, and for individual citations. *Voyant Tools* (Sinclair and Rockwell, 2016) is a collection of concordance tools and some of these also use visualisation to show the proximity of words with one another, or the relative frequency of words. Another approach is to obtain a statistical measure of the importance of a word, in relation to its occurrence within a text, using the metric ‘term frequency–inverse document frequency’ (*TF-IDF*).

A related, but distinct, approach is *automatic term recognition*, where a tool such as *Termine* combines statistical significance of words with a 'part of speech' parser to make linguistic associations from text (Frantzi et al., 2000). NaCTeM’s web demonstration tool of Termine presents terms and phrases as a ranked list based on its C-value (a statistical measure of the frequency and significance of term occurrence), and as an annotated text showing the terms that have been extracted by the tool (NaCTeM, 2016).

Automatic clustering analyses the distribution of terms (words) in small bodies of text (such as, titles and abstracts) and identifies groups of documents which use similar combinations of words; a descriptive term is applied to each cluster to aid human interpretation (Carpineto et al., 2009). We used the *Lingo3G* algorithm clustering utility from CarrotSearch.com (Carpineto et al., 2009), which is integrated within EPPI-Reviewer 4. It can generate clusters and hierarchical clusters or ‘subclusters within clusters’, depending on user preference. Citations may be present in one or more clusters, depending on the word combinations that are grouped together.

In the case study and examples below, the text mining tools used were generally open access with the exception of Endnote and Lingo3G. We utilised TF-IDF and Lingo3G automated clustering tool within the non-commercial subscription-based review management software program, EPPI-Reviewer 4.0 (Thomas et al., 2010). Termine was utilised through the NaCTeM website (NaCTeM, 2016).

Case study: using text mining tools and techniques for developing a search strategy

In the case study we compare the use of individual text mining tools and techniques to increase sensitivity through identifying suitable search terms, and to increase precision from examining preliminary outputs of a search for unwanted terms and concepts. The search strategy was developed to identify research literature on the social care and support of adults with intellectual disabilities as they get older. This was intended for a set of evidence reviews used to inform a NICE Guideline on the care and support of older people with learning disabilities (NICE, 2015). It was structured around broad terms for the population at individual and service level (older people, aged care) and health condition (intellectual disabilities, learning disabilities or named conditions). (The Medline search strategy is reproduced in Appendix 1.)

1. Increasing search sensitivity: comparison of TF-IDF, Termine and BibExcel

In order to identify suitable search terms to increase the sensitivity of the search, 52 study citations known to be of relevance to the topic area were analysed, collected from exploratory searches on the topic area. The quantity of the citations analysed was less important than the range of research collected. These citations were obtained from screening the results from a series of highly focused searches on areas considered relevant to the guideline, and were from PubMed and Applied Social Sciences Index and Abstracts (ASSIA) databases. TF-IDF values, Termine and BibExcel were used to analyse search terms and phrases in the titles and abstracts.

All the TF-IDF values were examined, consisting of 367 items (the value was of 6.5 or higher). Termine was used in conjunction with the part-of-speech parser (POS) Genie 2.1, which is customised to biomedical texts, and the first 60% of items in the ranked list was examined, representing 463 items. In BibExcel, words occurring in more than six citations were identified and their co-occurrence with another word was collected, which was an arbitrary cut-off point for ease of identifying any patterns in co-occurring words.

The resulting term lists were scanned for potentially relevant items relating to the population concept (older people, ageing) and the condition concept (intellectual disabilities). Suitable words identified from both the TF-IDF and Termine analyses were combined and used to search within the 52 items under analysis to determine how many citations would not be identified by these terms, and these citations were checked for potential search terms. Endnote was used to analyse the controlled vocabulary for the 52 citations.

A number of phrases for the population concept were identified from the TF-IDF analysis and Termine. As well as terms for older people, phrases relating to literature about ageing were identified, such as 'future planning', 'future care' and 'active ageing'. These terms were found to capture 44 of the 52 citations. Manually scanning the remaining eight citations led to further potential search terms being identified (longevity, aging adults, menopause, ageing factors, aging issues), which further informed the development of the search strategy. Four terms were identified from TF-IDF values and Termine that related to the health condition concept, and these located all but two records. These two records had no distinguishing health condition concept in the title and abstract; one mentioned intellectual disability in the controlled vocabulary field, and one mentioned intellectual disability in the journal title, which informed our strategy to search the journal name field. The final search strategy also included names of more health conditions.

The TF-IDF analysis and Termine yielded different results. Both produced a large ranked list of words, though the relative ranking of words differed. Table 1 shows examples of some of the significant words and their relative ranking based on the order in the generated word list. Table 1 also shows examples of some words not identified by Termine. The TF-IDF list comprised of a combination of single words and few phrases, and it contained the phrases 'older people', 'older adults', 'older person' and ‘menopause’ in the top 30 records. Terms relating to 'aged care' were much lower in the list. In comparison, the Termine list did not contain single words, and had phrases of at least two words. Some phrases describing older people, older adults were ignored by the algorithm, though ‘elderly people’ was ranked 72 in the first 463 phrases checked. However, Termine ranked ‘aged care’ much higher than TF-IDF, at 5 compared with 185. The Termine list included phrases that were not in the TF-IDF list, for example, 'late life', 'aging service', 'future living' and 'future perspective'. There was a difference in which word forms were used; for example, the TF-IDF list contained ‘menopause’, and twelve instances of 'menopause' in the sample were ignored by Termine, (though it identified the phrase 'menopause finding' lower than the 60% of phrases checked). However, Termine listed ‘menopausal’ at 168 (in a phrase 'carer menopausal attitude'), and this was not present in the TF-IDF list

For the health condition concept for 'intellectual disabilities', there were very few words from our sample. Both TF-IDF values and Termine revealed 'intellectual disability' and 'developmental disability'. Termine revealed two more conditions than TF-IDF: ‘down syndrome’ (ranked 21) and ‘learning disability’ (ranked 27).

The BibExcel list generated a list of single words. It showed that 'older' appeared in 35 out of 52 citations, and that ‘intellectual’ appeared in 42 out of 52 citations. The number of citations that contain at least one occurrence of the word, or two words is shown in Table 1. The BibExcel list was less helpful because most terms of interest were phrases, and single words were too generic, for example, ‘old’, ‘future’, ‘aged’.

2. Increasing search precision

While improving sensitivity helps to ensure that relevant literature is identified, improving *search precision* aims to minimise the identification of irrelevant literature. In this case, the search strategy was informed by text mining, but also integrated with search terms obtained from previous work, searches published in the literature, other NICE Guidelines, and iterative searching and browsing of citations from test searches within databases. Various iterations of the search strategy were run in Medline, however, the search appeared to be generating a large number of items irrelevant to the research question. To examine ways in which this could be minimised, and thereby increase precision of the search, 1,000 references from a test search were selected by date for three types of analysis: 1) Lingo 3G clustering to group the citations into labelled clusters; 2) BibExcel to obtain frequencies of citations that contained particular words; and 3) Endnote, to assess controlled terms at various iterations of the test searches.

In analysing the 1,000 records in Lingo3G, the 29 clustering labels generated did not reveal anything clearly that could be excluded (an extract is shown in Appendix 2). However, it facilitated examination of citations within the named clusters, such as 'Alzheimer's disease', to inform judgement on whether some concepts could be excluded from the search.

BibExcel was useful in producing a list showing how many words occurred within a citation. For example, 'gene' occurred in 10% of the citations and 'protein' and FMR1 each in 6% of citations. These approaches revealed a number of citations on genetic studies and studies concerning mental retardation protein, which were not of interest. Exploring the controlled terms in Endnote also informed further iterations of searches that were tried in order to remove some of the genetic studies. The final search used some exclusions for genetic studies within the controlled vocabulary for specific disease conditions, and the phrase 'mental retardation protein' was excluded where the phrase 'mental retardation' was used.

However, these steps alone did not sufficiently increase precision of the search output, though the text mining helped our realisation that the search contained many irrelevant results. Manual reflection was needed to reconsider the search strategy. The final search contained a conceptual modification to reduce the number of unwanted clinical studies, where the terms for specific conditions were required to be in close proximity to certain population or service user terms.

Further examples of applying tools in selected systematic reviews

In order to complement the findings of the case study, we reflect on the usefulness of some technologies beyond those identified in the case study above through some further examples.

Increasing precision of the search by combining multiple tools

When refining the search for a review of self-care and minor ailments (Richardson et al. *in press*), we needed to capture studies investigating primary care consultations, but reduce the number of irrelevant studies that would be found from only searching on the term 'primary care'. We analysed a sample of 54 records using the concordance tool AntConc which revealed the more precise phrases such as 'primary care practice', 'primary care consultancies', 'primary care centres', 'gp-supervised' and 'gp appointment'. The final search included these terms in close proximity with one another.

In a separate analysis for the same review, we analysed preliminary Medline search results that were limited to one publication year to investigate the presence of themes that were inadvertently captured by the search. Lingo 3G clustering of 428 items revealed a cluster 'cancer' (which was out of scope for the search), and we identified that this was being located owing to terms for ‘pain management’ and ‘pain control’. We next used some additional tools to examine the results of a subsequent search and analysis of 410 items from that year, which revealed other terms relating to ‘pain’. Using a function in Voyant Tools entitled 'document frequencies tool', we noticed that the word 'pain' appeared frequently in relation to 'chronic pain', 'chronic back' and 'chronic musculoskeletal'. An analysis of MeSH terms in Endnote revealed 'chronic disease' and 'chronic pain' in many items. BibExcel was used to discover that the word ‘pain’ was in nearly a quarter of the records, and the word ‘chronic’ was in a sixth of the sample. The final search was adjusted to reduce the number of unwanted items from chronic pain for certain conditions. For example, 'headache' was searched where present without the controlled term for chronic pain, or it was searched for without the freetext phrase 'chronic headache'. Voyant Tools' Cirrus word cloud tool revealed the presence of 'pandemic' in our sample, and although this was less predominant than the words 'chronic' and 'pain', it was noticeable from a brief check of the word cloud. As pandemics were an area outside the scope of the review, we could consider ways of limiting it within the search.

Increasing precision of the search through clustering

For a systematic review concerning relationships between exercise and osteoarthritis or chronic joint pain (Hurley et al., 2013), clustering was useful to aid in modifying the search of unwanted records. A sample of 3,655 items obtained from a draft PubMed search was clustered using Lingo 3G to assess dominant themes from the records located by the search strategy. This generated 29 clusters; cluster labels that were clearly recognisable as not within scope included 'total knee arthroplasty', 'total hip arthroplasty', and 'hip arthoscopy', which are types of surgical procedures, and 'rheumatoid arthritis'. By comparison, a TF-IDF analysis showed that the first mention of ‘arthroplasty’ was ranked 84th in the list of terms. Exploring the clusters led to the discovery that ‘arthroplasty’ was mentioned in nearly a fifth of citations from the test search. The final search was adjusted to reduce the number of unwanted items about surgery and post-operative recovery by excluding items containing surgery and post-operative recovery in their titles from the part of the search relating to osteoarthritis, and excluding surgery subheadings from the controlled vocabulary searches. We did not adjust the search terms to exclude for rheumatoid arthritis though some of these would have been reduced through reducing the number of citations on surgery.

In a different systematic review concerning medication errors in children (Sutcliffe et al., 2014), a test search in PubMed yielding 5,757 citations was clustered into groups using Lingo3G. This resulted in 28 clusters, and of these, several clusters were labelled with themes that were not included within the scope of the review: 'suicide attempts', 'pregnant women', 'illicit drugs', 'heroin overdose'. Citations were browsed within some of the clusters and term searches were also used to indicate how much literature there was on a topic on, for example, 'pharmacy', 'parents', 'suicide', 'traffic', and 'driving'. As a result, some elements were identified that could be potentially excluded from the search, with care not to exclude relevant items at the same time. These excluded elements related to: street drugs, alcohol behaviour, suicide, accidents while driving, and pregnancy.

It was previously reported that clustering was used on a set of records retrieved from a preliminary test search in PubMed relating to the late diagnosis of many health conditions (Stansfield et al., 2013). This identified a dominant theme that was not within the focus of the review (the genetic technique of polymerase chain reaction) and an amended search strategy that accounted for this theme reduced the number of records retrieved from the PubMed search by 4% (over 500 records). The amendment was also applied to databases searches of PsycINFO and CINAHL. A conservative estimate suggests that the additional PubMed records would have taken about half a day for one person to screen, representing a considerable workload saving (and having two people check records independently would double this estimate).

Discussion

We now summarise the potential utility and challenges in using these technologies for search term development. In particular, we: consider combining TF-IDF and Termine, compare word frequency and concordance tools; discuss the usefulness of clustering approaches; provide a brief description of visualisation and synonym tools; suggest sampling as a particular challenge in applying the tools; discuss considerations in using the tools; and consider potential limitations of this work.

The value of combining TF-IDF analysis and Termine

In the case study on the care and support of older people with intellectual disabilities, TF-IDF analysis and Termine were both found to be beneficial in identifying terms for one search concept (related to ageing). However, as each had differing results, their use is complementary to one another. Unlike TF-IDF, the Termine point of speech (POS) parsers are not intended to identify every word; in our example, the Genie 2.1 POS used here did not recognise phrases for older people or the term menopause. With the TF-IDF analysis, some terms of interest were either missed owing to parsing some phrases as single words, because the word was not significant in the body of text analysed. For both Termine and TF-IDF, the manual process of scanning the term lists has potential to miss items through a user either not recognising terms or phrase fragments as significant, or by not scanning the lower-ranked terms in the list. However, missing terms can be partly mitigated by iteratively using a technique of searching for citations not located by the search terms, and re-analysing successive citations.

O'Mara-Eves et al. (2014) observed a small challenge is deciding on the threshold below which terms identified by Termine would not be considered, and they used a threshold C-value of 5 when analysing the full text of five papers. Their rationale was that "it was the common value below which mined terms seemed to lose relevance across the five papers" (p. 53). Such thresholds are ultimately subjective and cannot be standardised across reviews, as terms with lower rankings might be relevant in some instances because the distribution of C-value scores returned will differ from corpus to corpus. For the review about self-care of minor ailments (Richardson et al. *at peer review*), Termine was applied to 51 title and abstract citations, initially using a threshold C-value of 5, which identified 22 citations; but then relevant terms that located more records were identified by those with a C-value between 2 and 5, identifying 40 out of 50 citations. With this technique, the C-value threshold used is less important as this is not the sole method for generating search terms.

Comparison of word frequency and concordance tools

BibExcel was particularly useful for obtaining the frequency of citations that contain particular words that were indicative of citations that were not of interest. A tool to analyse phrases giving the frequency per citation would have been better, had we wanted to consider this further. This is possible with AntConc, which has much more functionality than BibExcel for analysing words in text. We have shown that AntConc can provide more informative analysis of assessing phrases and co-located words within a specified distance of each other.

The usefulness of clustering approaches

Clustering can generate groups of citations rapidly. It draws upon the most dominant themes depending on the uniformity of the discriminating terms. Clustering is not as useful in a body of literature where the terms are interconnected or there is no dominant vocabulary to express a collection of unwanted (or wanted) items, as shown in the case study on older people with intellectual disabilities. However, in the other examples described, there were clear dominant themes that were unrelated to the area of interest, and these could be identified and attempts made to address this in the search strategy. From these experiences, we conclude that it is difficult to predict in advance when clustering might be useful.

We previously observed that two hierarchical tiers of clusters to be better than one tier for exploring themes in the dataset, as this provided better differentiation of topics. (Stansfield et al., 2013). However, in obtaining an overview of the literature for assessing the performance of the search strategy, it can be useful to have both single and two tiers of clusters to explore, as the single tiered clusters allow the collection of citations within an overarching cluster label to be observed.

Within a clustering algorithm, there are two separate processes. The main process is core clustering, which is typically a mathematical analysis of the distribution of terms. The second process is finding a good label to describe the clusters. The Lingo3G algorithm has addressed both these aspects, but when evaluating the utility of a given clustering solution it is important to bear in mind that both are being evaluated at once. Other clustering algorithms such as Latent Dirichlet Allocation may identify more coherent clusters than Lingo 3G (i.e. are better at finding the similarities between groups of citations), but as they do not identify simple labels – but offer ordered lists of terms – significant user interpretation is needed to identify *why* particular groups of citations have been put together (Carpineto et al., 2009).

Visualisation tools

It is possible to link the output of text mining to visualisation tools, such as word clouds. While this may provide a quick overview and have visual appeal, it is unclear how these could offer more meaningful information than a ranked list of terms or phrases. For example, when words are presented at different angles and in a range of colours, it could be easy to miss some important words. However, this might improve with development and integration with other tools. The Cirrus word cloud tool allows user control of the number of highest frequency words that are displayed, offering some flexibility of appearance. Some visualisation tools, such as VOS-Viewer, can show keyword co-occurrence networks, where the distance between two terms provides an indication of the number of co-occurrences (van Eck and Waltman, 2016). As such, it may reveal possible areas of citations containing unwanted items in a search. For example, Glanville (2016) showed the word, 'recruitment' was present in a search sample in conjunction with the separate concepts of clinical trials and molecular biology.

Synonym tools

Distinct from these tools are other tools that rely on an external corpus of literature to provide relevant terms (by ‘external corpus’, we mean a corpus of studies outside of the review). A noticeable absence from the literature concerns tools that identify synonyms or homonyms, particularly outside the medical literature; such tools rely on an external corpus. For example, NaCTeM 's History of Medicine semantic search system includes synonyms and a range of other semantically related medical terms (Thompson et al., 2016), drawing on two archives of historical medical text. Such tools have the potential to provide a more objective perspective of appropriate search terms within an area, beyond those obtained from a user-derived sample.

Sampling as a particular challenge in applying the tools

A key challenge is using a suitable sample of studies to analyse. Careful consideration is needed to avoid introducing selection bias. If the purpose of text mining is either to increase sensitivity, or ensure the search is of a good standard, there is potential for this process to instigate a situation whereby the sample used may only reveal more of the same, or what one expects to be there, because of the way the sample was collected in the first place. In the case study and examples described here, the samples of studies were collected to increase sensitivity for selected topics that were difficult to describe, and they were not intended to identify all of the search terms. The quantity of citations collected for the sample was arbitrary, though the samples were intended to comprise of a range of relevant concepts. If text mining is being used to refine a search, perhaps to reveal unwanted items in a collection of research for the purpose of increasing precision, the sample might simply be the citations (or subset, or specific timeframe of citations) from a test search strategy or a search line within a search strategy.

The data included in sample are also important to consider. In the examples here, citations and abstracts were used; however, O'Mara-Eves et al. (2014) used the full-text of five papers that were seminal within the area of their search focus. In some cases, the use of full-texts may not always be possible (for example, limitations in the software, or a lack of known relevant studies), or it might be too inefficient to make the process worthwhile, given that retrieving and then the processing the full-text documents may add considerable time. Whether better quality information can be gleaned from abstracts versus full-texts is unknown, but is dependent on the breadth and depth of relevant terms used in each, which cannot be known in advance and will likely vary from citation to citation, and review to review.

Considerations in using the tools

Where text mining is applied to a collection of citations and abstracts, it is particularly useful to understand how many citations relate to a given term, in order to indicate the relative impact of a term in locating the items from a search. Without knowing how many terms a given citation is responsible for generating, the terms from long citations or documents with repeating words may be over-represented within a sample and appear high in a ranked list, even though they might only relate to one citation. In utilising the above tools to improve precision, rapidly generating this information from BibExcel and Endnote helped inform whether it was worth spending time exploring specific words in the search.

The text mining tools and techniques used here were applied quickly, with the bulk of the time spent on analysing the results. The iterative process of developing the search strategies took time as they had particular challenges in achieving a search that balanced sensitivity and precision. Using text mining tools alongside other methods for search term development requires additional time input, though it is difficult to quantify how much additional effort is involved. Partly this depends on the complexity of the search task and the extent of text mining undertaken. Related to this is judging how best to utilise the tools for a particular purpose, and when to stop developing the search. The time required for both will vary depending on the familiarity of the user with the tools and approaches employed. However, the additional time has potential to improve the quality of the search and potentially reduce the volume of records retrieved – thus saving time further into the review process.

For most tools used, an element of pre-processing of the citations within a reference management tool was needed in order to analyse specific citation fields (such as titles, abstracts, keywords); though this was not onerous, some familiarisation with the process was needed. If we wanted to find out how many citations in a sample a term related to, some tools (for example, Termine) required multiple steps, such as combining the tool with citation or review management software to 'search within' to obtain the relevant citations

The process of generating terms from a prepared sample using Termine or TF-IDF analysis takes less than a minute and little learning time on its use is required. A drawback of AntConc is that additional processing is required to separate records into citations (Hausner et al., 2011), accessibility can be hampered by institutional firewalls, and some learning on how to use the tool is necessary. In comparison, BibExcel was quicker to apply, although more time was initially needed to understand and develop the steps to utilise BibExcel because it is multi-functional tool (Gourlay (2010) was a helpful starting point). Also, BibExcel has less functionality as a concordance tool. The Voyant Tools were accessible and rapid to use through their web-interface, though some functions do not facilitate understanding how many citations relate to a term, without using citation or review management software to 'search within' for that term.

In terms of costs, it would be useful to undertake a full economic analysis of this part of the systematic review workflow, though careful prospective design would be needed to capture the necessary information. Shemilt et al. (2016) undertook a cost-effectiveness analysis of the title-abstract screening stage of a systematic review, but did not examine the impact of different methodologies for constructing the initial search. Unlike the screening stage of the review, search strategy design is an iterative process, influenced by human knowledge and skills, and may be approached in a variety of ways depending on the purpose and resources of a review; these issues would need to be carefully considered in an economic evaluation.

Potential limitations of this work

We have not investigated how the use of these tools might interact with, or be supplanted by, emerging tools and methodologies for using machine learning in the citation screening process. For example, we have described above how tools can be used iteratively to improve the precision of a search by identifying terms and concepts which lie outside the scope of a review. It may be, however, that the process of 'active learning', whereby the machine is able to 'learn' to distinguish between relevance and irrelevance would result in a machine learning model that identifies irrelevant terms automatically; thus, the incremental time saved in reducing the search yield using the methods described in this paper may be reduced. However, this would probably mean accepting machine judgements for excluding some citations without any manual checking – something which may require more empirical evidence before it be adopted widely (O'Mara-Eves et al., 2015; Thomas, 2013). It also means accepting uncertainties on the potential size of the literature that needs to be screened manually. In the case study, the search was developed for multiple questions and, because the screening was undertaken manually, there was a need to tailor the search to the resources available to screen.

Finally, the case study set out to compare the usefulness of different tools in aiding search term identification, and increased understanding of the advantages and limitations of their use. The other examples presented were selected to show where text mining tools have been useful, as they were documented as part of the process of developing the search strategies in the reviews concerned. However, this does not give a comprehensive picture of situations where the use of these tools was *not* useful, as such instances were not documented. From the examples here, it seems difficult to predict how useful text mining may be for individual search strategies, particularly for diverse literature, owing to the nature of language and the potential studies of interest. Nonetheless, by using fairly rapid, easily available tools, text mining is likely to be an appealing approach to complement other search processes.

Conclusion

This paper identifies five applications of text mining for search term development: increasing sensitivity and increasing precision of searches, aiding translation of searches across databases, searching and screening within an integrated repository system, and developing objective search strategies. Using a case study and further examples, the paper explores the usefulness and challenges of using some text mining tools for two applications: increasing sensitivity and precision. We found that text mining can aid the discovery of search terms for search strategies for diversely-described topics to support an iterative search strategy development process. Using multiple tools appears to be particularly fruitful. Their usefulness is influenced by the varying functionality of the tools used, the way that they are used, and the text that is analysed. An awareness of how the tools perform can help utilise them more efficiently and effectively, though the overriding challenge of finding efficient ways to identify an unknown body of literature for incorporation in systematic reviews still remains.

Acknowledgements

CS undertook the review and main analysis. CS, AO and JT reflected on the findings and participated in writing the manuscript. JT directs the development and management of EPPI-Reviewer, the EPPI-Centre's software for systematic reviews. Thank-you to Jenny Caird for feedback and support on early drafts of this paper.

This work was undertaken by the EPPI-Centre, which received funding from the Department of Health. The views expressed in the publication are those of the authors and not necessarily those of the Department of Health. The case study includes work undertaken as part of the work of the NICE Collaborating Centre for Social Care which received funding from the National Institute for Health and Care Excellence. The views expressed in this publication are those of the authors and not necessarily those of the Institute.

References

Anthony L. 2014. AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Damarell RA, Tieman JJ, Sladek RM. 2013. OvidSP Medline-to-PubMed search filter translation: a methodology for extending search filter range to include PubMed's unique content. *BMC Medical Research Methodology* **13**: 86.

EUnetHTA. 2015. *Guideline: Process of information retrieval for systematic reviews and health technology assessments on clinical effectiveness.* http://www.eunethta.eu/outputs/eunethta-methodological-guideline-process-information-retrieval-systematic-reviews-and-healt [accessed 12.8.2016]

Frantzi, K., Ananiadou, S. and Mima, H. 2000. Automatic recognition of multi-word terms. *International Journal of Digital Libraries* **3**(2): 117-132.

Glanville J. 2016. Text mining for strategy development. Presentation at NICE Joint Information Day, 8 March 2016 London.

Goss C, Lowenstein S, Roberts I, DiGuiseppi C. 2007. Identifying controlled studies of alcohol-impaired driving prevention: designing an effective search strategy. *Journal of Information Science* 33: 151-162.

Gourlay S. 2010. Preparing to review the literature systematically. v.3.12. Kingston Business School.

Hausner E, Glanville J, Waffenschmidt S. 2011. Workshop 'Text analysis tools for information retrieval', 22.10.2011 19th Cochrane Colloquium Madrid.

Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. 2015. Development of search strategies for systematic reviews: validation showed the noninferiority of the objective approach. Journal of Clinical Epidemiology 68: 191-199.

Hausner E, Guddat C, Hermanns T, Lampert U, Waffenschmidt S. 2016. Prospective comparison of search strategies for systematic reviews: an objective approach yielded higher sensitivity than a conceptual one. *Journal of Clinical Epidemiology* doi: 10.1016/j.jclinepi.2016.05.002.

Hausner E, Waffenschmidt S, Kaiser T, Simon M. 2012. Routine development of objectively derived search strategies. *Systematic Reviews* 1.

Hayman S, Shaheem Y. 2014. Smart Searching: Logical Steps to Building and Testing Your Literature Search. CareSearch Palliative Care Knowledge Network. Date retrieved: 3 January 2017 URL: http://sites.google.com/site/smartsearchinglogical/home

HLWIKI Canada contributors. 2016. *PubMed Alternative Interfaces HLWIKI Canada*, Date of last revision: 18 May 2016; Date retrieved: 11 August 2016 URL: http://hlwiki.slais.ubc.ca/index.php?title=PubMed_Alternative_Interfaces&oldid=144597 Page Version ID: 144597

Hurley M, Dickson K, Walsh N, Hauari H, Grant R, Cumming J, Oliver S. 2013. Exercise interventions and patient beliefs for people with chronic hip and knee pain: a mixed methods review (Protocol). *Cochrane Database of Systematic Reviews,* Issue 12. *DOI: 10.1002/14651858.CD010842.*

Kok R, Verbeek JA, Faber B, van Dijk FJ, Hoving JL. 2015. A search strategy to identify studies on the prognosis of work disability: a diagnostic test framework. *BMJ Open* **5**: e006315.

Lunny C, McKenzie JE, McDonald S. 2016. Retrieval of overviews of systematic reviews in MEDLINE was improved by the development of an objectively derived and validated search strategy. *Journal of Clinical Epidemiology* **74**: 107-118.

Mergel GD, Silveira MS, da Silva TS. 2015. A method to support search string building in systematic literature reviews through visual text mining. SAC '15 Proceedings of the 30th Annual ACM Symposium on Applied Computing 1594-1601.

NaCTeM. 2016. *Termine Web Demonstrator* http://www.nactem.ac.uk/software/termine/ [accessed 12.8.2016]

NICE. 2015. *NICE guideline: Care and support of older people with learning disabilities – Final Scope*, *17-12-2015* https://www.nice.org.uk/guidance/GID-SCWAVE0776/documents/final-scope-2

O'Mara-Eves A, Brunton G, McDaid D, Kavanagh J, Oliver S, Thomas J. 2014. Techniques for identifying cross-disciplinary and hard-to-detect evidence for systematic review. *Research Synthesis Methods* **5**: 50-59.

O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic Reviews* 4: 5.Paynter R, Bañez LL, Berliner E, Erinoff E, Lege-Matsuura J, Potter S, Uhl S. 2016. *EPC Methods: An Exploration of the Use of Text-Mining Software in Systematic Reviews* [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US)

Perrson O. Undated. *BibExcel. A tool-box developed by Olle Persson.* http://homepage.univie.ac.at/juan.gorraiz/BibExcel/ [accessed 26.8.2016]

Petrova M, Sutcliffe P, Fulford KWM, Dale J. 2012. Search terms and a validated brief search filter to retrieve publications on health-related values in Medline: a word frequency analysis study. *Journal of the American Medical Informatics Association* 19: 479-488.

Richardson M, Khouja C, Sutcliffe K, Hinds K, Stansfield C, Thomas J (in press) *Decision-making and behaviour change in self-care for minor ailments: a systematic review to evaluate how people decide, and how they can be directed, to use the most appropriate service*. London: EPPI Centre, Social Science Research Unit, UCL Institute of Education, University College London.

Shemilt I, Khan N, Park S, Thomas J. 2016. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *Systematic Reviews* **5**:140.

Simon M, Hausner E, Klaus S, Dunton N. 2010. Identifying nurse staffing research in Medline: development and testing of empirically derived search strategies with the PubMed interface. *BMC Medical Research Methodology* 10.

Sinclair S, Rockwell G. 2016. *Voyant Tools Web*. http://voyant-tools.org/.

Stansfield C, Thomas J, Kavanagh J. 2013. Clustering documents automatically to support scoping reviews of research: a case study. *Research Synthesis Methods* **4**: 230-241.

Sutcliffe K, Stokes G, O'Mara-Eves A, Caird J, Hinds K, Bangpan M, Kavanagh J, Dickson K, Stansfield C, Hargreaves K, Thomas J. 2014. *Paediatric medication error: a systematic review of the extent and nature of the problem in the UK and international interventions to* *address it*. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London. ISBN: 978-1-907345-73-9

Thomas J. 2013. Diffusion of innovation in systematic review methodology: Why is study selection not yet assisted by automation? *OA Evidence-Based Medicine* **1**:1–6.

Thomas J, Brunton, J, Graziosi S. 2010. *EPPI-Reviewer 4.0: Software for Research Synthesis. EPPI-Centre Software*. London: Social Science Research Unit, Institute of Education, University of London.

Thomas J, McNaught J, Ananiadou S. 2011. Applications of text mining within systematic reviews. *Research Synthesis Methods* **2**: 1–14.

Thompson J, Davis J, Mazerolle L. 2014. A systematic method for search term selection in systematic reviews. *Research Synthesis Methods* **5**: 87-97.

Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E, McNaught J, Timmermann C, Worboys M, Ananiadou S. 2016. Text Mining the History of Medicine. *PLoS ONE* **11**:1, e0144717.

van Eck NJ, Waltman L. 2016. *VOSviewerManual. Manual for VOSviewer version 1.6.4*. Leiden University.

Wentz R, Roberts A, Bunnv F, Edwards P, Kwan I, Lefebvre C. 2001. Identifying controlled evaluation studies of road safety interventions: searching for needles in a haystack. *Journal of Safety Research* **32**: 267-276.

White VJ, Glanville J, Lefebvre C, Sheldon TA. 2001. A statistical approach to designing search filters to find systematic reviews: objectivity enhances accuracy. *Journal of Information Science* **27**: 357-370.

Zhang H, Babar MA, Tell P. 2011. Identifying relevant studies in software engineering. *Information and Software Technology* **53**: 625-637.

Table 1: Comparison of selected words from text mining 52 citations from different tools

Phrase	**TF-IDF relative rank** (1 = highest rank)	**Termine relative rank** (1 = highest rank)	**BibExcel Number of citations –word co-occurrence** (minimum threshold= 6)
Older people	7	n/a	24
Older adult	20	n/a	Older adults = 18
Older person	25	n/a	6
Menopause	10 (menopausal not listed)	'menopausal' at 168	n/a
Retirement	55	72	13
Active ageing	53	18	n/a
Aged care	185	5	9
Community based aged care	128	15	n/a
Future care	281	72	9
Future planning	74	21	n/a