Journal of Biomedical Informatics 51 (2014) 242–253

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Reducing systematic review workload through certainty-based screening

Makoto Miwa [a,b,*,1], James Thomas [c], Alison O'Mara-Eves [c], Sophia Ananiadou [a]

[a] The National Centre for Text Mining and School of Computer Science, Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester M1 7DN, UK
[b] Toyota Technological Institute, 2-12-1 Hisakata, Tempaku-ku, Nagoya 468-8511, Japan
[c] Evidence for Policy and Practice Information and Coordinating (EPPI-)Centre, Social Science Research Unit, Institute of Education, University of London, London, UK

ARTICLE INFO

Article history:
Received 10 December 2013
Accepted 7 June 2014
Available online 19 June 2014

Keywords:
Systematic reviews
Text mining
Certainty
Active learning

ABSTRACT

In systematic reviews, the growing number of published studies imposes a significant screening workload on reviewers. Active learning is a promising approach to reduce the workload by automating some of the screening decisions, but it has been evaluated for a limited number of disciplines. The suitability of applying active learning to complex topics in disciplines such as social science has not been studied, and the selection of useful criteria and enhancements to address the data imbalance problem in systematic reviews remains an open problem. We applied active learning with two criteria (certainty and uncertainty) and several enhancements in both clinical medicine and social science (specifically, public health) areas, and compared the results in both. The results show that the certainty criterion is useful for finding relevant documents, and weighting positive instances is promising to overcome the data imbalance problem in both data sets. Latent dirichlet allocation (LDA) is also shown to be promising when little manually-assigned information is available. Active learning is effective in complex topics, although its efficiency is limited due to the difficulties in text classification. The most promising criterion and weighting method are the same regardless of the review topic, and unsupervised techniques like LDA have a possibility to boost the performance of active learning without manual annotation.

1. Introduction

Systematic reviews are a widely used method to bring together the findings from multiple studies in a reliable way and are often used to inform policy and practice (such as guideline development). A critical feature of a systematic review is the application of the scientific method to uncover and minimise bias and error in the selection and treatment of studies [1,2].

As a result, reviewers make efforts to identify all relevant research for inclusion in systematic reviews. However, the large and growing number of published studies, and their increasing rate of publication, makes the task of identifying relevant studies in an unbiased way both complex and time consuming. Moreover, the specificity of sensitive electronic searches of bibliographic databases is low. In a process known as screening, reviewers often need to look manually through many thousands of irrelevant titles and abstracts in order to identify the much smaller number of relevant ones [3]. Reviews that address complex health issues or that deal with a range of interventions are often those that have the most challenging numbers of items to screen. Given that an experienced reviewer can take between 30 s and several minutes to evaluate a citation [4], the work involved in screening 10,000 citations is considerable (and the screening burden in some reviews is considerably higher than this).

Text mining facilitates the reduction in workload in conducting systematic reviews in a range of areas [5–7]. Text mining is used increasingly to support knowledge discovery, hypothesis generation [8] and to manage the mass of literature. Its primary goal is to extract new information such as relations hidden in text between named entities and to enable users to systematically and efficiently discover, collect, interpret and curate knowledge required for research [9]. The technology most often tested in relation to the reduction in screening burden is automatic classification, where a machine 'learns', based on manual screening, how to apply inclusion and exclusion criteria [10]; that is, it semi-automates the screening process. Pertinent to the focus of this paper, there have been a range of evaluations of the

* Corresponding author at: Toyota Technological Institute, 2-12-1 Hisakata, Tempaku-ku, Nagoya 468-8511, Japan. Tel.: +81 (0)52 809 1760.
E-mail addresses: makoto-miwa@toyota-ti.ac.jp (M. Miwa), j.thomas@ioe.ac.uk (J. Thomas), a.omara-eves@ioe.ac.uk (A. O'Mara-Eves), Sophia.Ananiadou@manchester.ac.uk (S. Ananiadou).
[1] This work was carried out while the author was at the University of Manchester, Manchester, UK.

http://dx.doi.org/10.1016/j.jbi.2014.06.005
1532-0464/

performance of various text mining tools to reducing screening burden, some of which have achieved reductions in workload of between 50% [4] and 90–95% [11,12] (though others have had rather less success [13]).

The nature of the contribution that such methods can make to systematic reviews is the subject of ongoing debate and evaluation. In some contexts, every citation needs to be screened by two reviewers, and in such situations the workload reduction applies only to the "second" reviewer, with all citations being screened by a human: the theory being that this will maximise recall [14]. In other contexts, citations are checked by a single reviewer, and the theory behind semi-automation is that some of these citations need not be screened manually; here, acceptable recall values are high, in the 95–99% range, but do not necessarily require 100% recall [15]. In a third context, automation is used simply to prioritise workload and ensure that the most likely relevant citations are screened earlier on in the process than would otherwise be the case [12]. Whichever situation pertains, there is a need to optimise the performance of the (semi-) automation methods used in order to maximise both recall and precision (see [13]).

While some studies have yielded impressive results, we lack instances in diverse contexts. In particular, most previous work has been undertaken in systematic reviews of clinical interventions, and the literature in this area is likely to have distinct advantages for machine learning which might not apply universally. Firstly, the use of technical terminology is widespread, and specific terms (e.g., drug names, proteins, etc.) are used in precise ways in distinct literature, in contrast to some disciplines where complex and compound concepts may be used (e.g., 'healthy eating' can be described in many ways). Secondly, the medical literature is well indexed on major databases (notably MEDLINE), with the availability of manually assigned Medical Subject Heading (MeSH) terms affording additional information to a classifier; such information is not present on the citations downloaded from other databases. There is therefore a need to assess the performance of text mining for screening in systematic reviews of complex, non-clinical contexts where the use of controlled vocabularies is variable or non-existent.

One of the main strategies adopted in previous work with automatic classifiers is active learning [4]. This 'supervised' machine learning technique involves beginning with a small training set and, through iteration, the training set is increased in size and utility (see Fig. 1). Once a given stopping criterion is reached (for example, when all relevant studies have been identified, or when the reviewers have run out of time for manual screening), the process ceases, and the remainder of studies not yet screened manually is discarded. There is thus a good 'fit' between the screening process in a systematic review, and the method of active learning. As manual screening progresses, the quantity of training material increases, and there is the opportunity for the classifier to 'suggest' items for manual screening, thus making the process more efficient. Although there is an accepted risk when automation is used that some relevant studies may be missed, the gains in terms of reducing burden might make this approach worthwhile. An evaluation of the trade-off between potentially missing studies and reducing burden is required. Given the concerns raised about using such technologies in complex topics, it is important to evaluate performance over a range of conditions.

The primary aim of this study is to assess the suitability of active learning applications to screening in systematic reviews of complex topics, with an emphasis on determining optimal conditions for running these technologies. This paper therefore addresses the following research questions:

1. Does active learning demonstrate similar performance (reduction of burden) in systematic reviews of public health (complex topics) as observed in clinical areas?
2. What features of the active learner improve performance? Specifically,
   (a) does the criterion used to determine the next instances to be annotated in the active learning cycle (i.e., certainty or uncertainty) affect performance? And
   (b) do different types of enhancements to the classifier affect performance?

2. Methods

Active learning methods can be classified into two categories from the perspective of data processing: pool-based and stream-based [16]. Pool-based active learning methods assume an unlabelled pool of instances, and determine the most appropriate instances to be annotated from a given data set by sorting them in terms of their informativeness. They often require considerable computational cost and memory. In contrast, stream-based active learning methods receive instances one at a time and decide whether or not the instance should be annotated. However, experiments suggest that stream-based approaches can have poor learner rates and raise too many unnecessary queries compared to pool-based approaches [17].

In this paper, we focus on pool-based active learning methods, since we are interested in learning from specific data sets, in which the sparse positive instances should be identified and presented for annotation as early as possible during the annotation process.

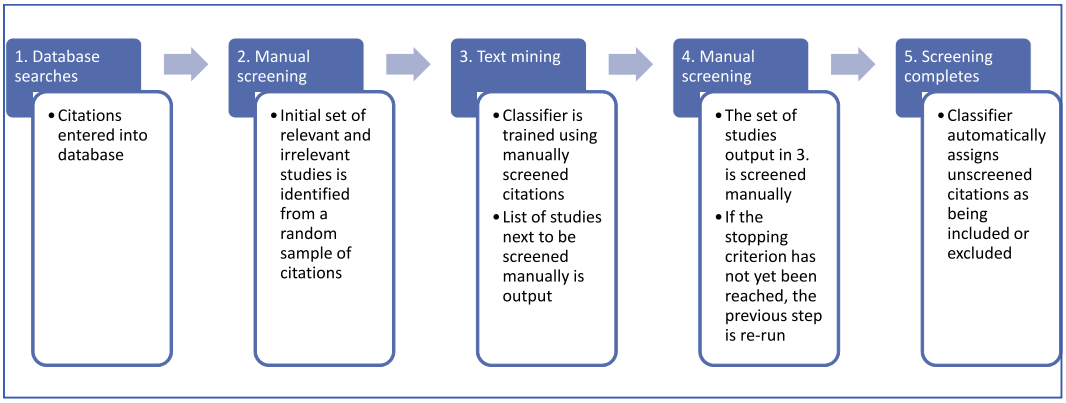

Fig. 1. The active learning process.