Journal of Biomedical Informatics 72 (2017) 67–76

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

A semi-supervised approach using label propagation to support citation screening

Georgios Kontonatsios [a], Austin J. Brockmeier [b], Piotr Przybyła [a], John McNaught [a], Tingting Mu [a], John Y. Goulermas [b], Sophia Ananiadou [a,*]

[a] National Centre for Text Mining, School of Computer Science, University of Manchester, Manchester, United Kingdom
[b] School of Electrical Engineering, Electronics and Computer Science, University of Liverpool, Liverpool, United Kingdom

ARTICLE INFO

Article history:
Received 15 February 2017
Revised 30 May 2017
Accepted 21 June 2017
Available online 23 June 2017

Keywords:
Active learning
Label propagation
Citation screening
Semi-supervised learning
Text classification

ABSTRACT

Citation screening, an integral process within systematic reviews that identifies citations relevant to the underlying research question, is a time-consuming and resource-intensive task. During the screening task, analysts manually assign a label to each citation, to designate whether a citation is eligible for inclusion in the review. Recently, several studies have explored the use of active learning in text classification to reduce the human workload involved in the screening task. However, existing approaches require a significant amount of manually labelled citations for the text classification to achieve a robust performance. In this paper, we propose a semi-supervised method that identifies relevant citations as early as possible in the screening process by exploiting the pairwise similarities between labelled and unlabelled citations to improve the classification performance without additional manual labelling effort. Our approach is based on the hypothesis that similar citations share the same label (e.g., if one citation should be included, then other similar citations should be included also). To calculate the similarity between labelled and unlabelled citations we investigate two different feature spaces, namely a bag-of-words and a spectral embedding based on the bag-of-words. The semi-supervised method propagates the classification codes of manually labelled citations to neighbouring unlabelled citations in the feature space. The automatically labelled citations are combined with the manually labelled citations to form an augmented training set. For evaluation purposes, we apply our method to reviews from clinical and public health. The results show that our semi-supervised method with label propagation achieves statistically significant improvements over two state-of-the-art active learning approaches across both clinical and public health reviews.

1. Introduction

Systematic reviews are used to identify relevant citations and answer research questions by gathering, filtering, and synthesising research evidence. A primary objective of any systematic review is to minimise publication bias [1] by analysing all citations relevant to the review. To identify and subsequently analyse every possible eligible study, reviewers need to exhaustively filter out citations (retrieved by searches to literature databases) that do not fulfill the underlying eligibility criteria. Developing systematic reviews is a time-consuming and resource intensive process that can take more than a year, with up to half of this time being spent searching and screening hits. As an example, an experienced reviewer requires 30 s on average to decide whether a single citation is eligible for inclusion in the review, although this can extend to several minutes for complex topics [2]. This amounts to a considerable human workload, given that a typical screening task involves manually screening thousands of citations [3–5].

To reduce the time and cost needed to complete the screening phase of a systematic review, researchers have explored various techniques, including crowdsourcing and text mining methods. Crowdsourcing approaches efficiently address tedious tasks, e.g., assessing the quality of Wikipedia articles [6], by re-distributing the overall workload to a large network of people. In the context of systematic reviews, the EMBASE screening project,[1] a Cochrane initiative, adopts a crowdsourcing approach to identify reports of randomised controlled trials (RCTs) and quasi-RCTs in the EMBASE bibliographic database. Two years after the project started, 4606

* Corresponding author.
E-mail address: sophia.ananiadou@manchester.ac.uk (S. Ananiadou).

[1] http://www.researchgate.net/project/The-Embase-project.

http://dx.doi.org/10.1016/j.jbi.2017.06.018
1532-0464/

crowd workers have processed a total number of 1 million EMBASE abstracts. Regarding the quality of the screening decisions, the crowd workers were found to be very accurate achieving a sensitivity and specificity performance of 99%.

In addition to crowdsourcing approaches, previous studies have investigated the use of automatic text classification to facilitate citation screening of systematic reviews [5,7]. In citation screening supported by automatic text classification, a human reviewer needs to screen only a subset of the retrieved citations. The process starts with a subset of citations manually annotated with labels, which denote whether the citation should be included or excluded. The citations paired with the labels serve as the training examples for the automatic classifier. In a supervised learning manner, the classifier is then trained on the manually annotated set to learn how to discriminate between relevant and irrelevant citations. As a final step, the trained classifier is applied to automatically screen the remaining unlabelled citations.

In this paper, we focus on a special case of automatic text classification known as feedback-based or active learning classification [2,8–11]. Active learning classification approaches are supervised machine learning methods that are iteratively trained on an increasing number of manually labelled citations. At each learning cycle, the model selects a small sample of citations and interactively requests a human analyst to manually label the citations. The manually labelled sample of citations is added to the training set and the model is retrained (updated). Results obtained by previous work [2,8] demonstrate that active learning classification approaches can substantially decrease the screening workload without reducing the sensitivity of the review (i.e., the method identifies 95–100% of relevant citations). However, a drawback of existing active learning methods is that the underlying model yields a low performance during the early learning iterations due to the limited number of labelled citations used as training instances. This can be explained because active learning methods exploit machine learning models whose hypothesis space, i.e., the possible set of decision boundaries, is constrained by the number training instances. Thus, a small number of training samples in the initial stages may result in poor classification performance [12].

Previous work [5,13] has outlined that the early identification of eligible citations presents several advantages to systematic reviewers and can significantly accelerate the overall citation screening process. As an example, O'Mara-Eves et al. [5] argued that, in a manually conducted citation screening task, reviewers tend to screen at a lower rate during the initial stages of the task while they incrementally increase their screening rate only after processing a larger number of eligible citations. Thus, the prioritisation of eligible citations during the initial active learning iterations can enable reviewers to establish a higher screening rate early in the process, reducing in this way the overall time needed to complete the citation screening task.

Based upon this, we propose a semi-supervised active learning method to improve the classification performance of active learning during the early stages of screening. In our approach, we adopt the 'cluster assumption' [14], which states that instances that are similar to each other will often share the same label. Accordingly, we use label propagation [15] to copy the label from a manually labelled citation to similar unlabelled citations (which are nearby in the feature space). These pseudo-labelled samples are used as additional training data for the classifier. To compute pairwise similarities between labelled and unlabelled instances, we explore two different feature representations of citations: (a) a bag-of-words feature space which consists of words that occur in the title and/or in the abstract of the citation and (b) a spectral embedding space that approximates the similarities of the bag-of-words representation based on their relative location in a lower dimensional space (neighbouring instances in the embedding should share similar content).

The label propagation step, which extends the training set with additional pseudo-labelled instances, can be used with any active learning method. Here, we integrate the proposed label propagation method with two existing active learning strategies, namely a certainty-based [8] and an uncertainty-based active learner [2]. The two strategies have different motivations. Uncertainty-based sampling [16,11] learns to discriminate between eligible and ineligible citations by requesting feedback from an analyst on citations that are more likely to change the current model. Certainty-based sampling [8,17] seeks to identify the relevant citations as early as possible, which is a useful strategy for systematic reviews [5].

For experimentation, we investigate the performance of the semi-supervised active learning method when applied to both clinical and public health systematic reviews. Such reviews are becoming increasingly difficult to manually develop and update due to the exponential growth of the biomedical literature (e.g., on average 75 trials and 11 systematic reviews are published daily in MEDLINE [18]). As an example, only a third of systematic reviews in the Cochrane library are being frequently updated with new relevant evidence published in the literature [19]. Thus, semi-automatic methods that can potentially accelerate the development of clinical and public health reviews are needed [20].

The contributions that we make in this paper can be summarised in the following points: (a) we propose a new semi-supervised active learning method to facilitate citation screening in clinical and public health reviews; (b) we show that a low-dimensional spectral embedded feature space can more efficiently address the high terminological variation in public health reviews versus the bag-of-words representation; and (c) experiments across two clinical and four public health reviews demonstrate that our method achieves significant improvements over two existing state-of-the-art active learning methods when a limited number of labelled instances is available for training.

1.1. Previous work on automatic citation screening

Previous approaches to automatic citation screening can be coarsely classified into automatic text classification and active learning classification methods. Aphinyanaphongs and Aliferis [21] proposed one of the earliest automatic text classification approaches for identifying high-quality and content-specific research articles useful for evidence-based reviews. They experimented with different supervised machine learning methods including a naïve Bayes classifier [22], boosting [23] and a support vector machine (SVM) [24]. As the feature representation for articles, they exploited words occurring in the title and/or in the abstract, the publication type (e.g., randomised control trial) and MeSH terms. Experimental results determined that the SVM classifier achieved an improved classification performance over the naïve Bayes and boosting classifiers.

Cohen et al. [13] applied an automatic text classification model in 15 systematic reviews relating to drug class efficacy for disease treatment. They used a modified version of the voted perceptron algorithm [25], i.e., a maximal-margin classifier which, similarly to an SVM, tries to find a hyperplane to separate relevant from irrelevant citations. As in previous work [21], they used a bag-of-words feature representation complemented by publication type and MeSH term features. In order to better address the high-recall requirement of systematic reviews—that is, reviewers need to identify all relevant citations for inclusion in the review—they introduced a bias weight to control the learning rate of positive (relevant) and negative (irrelevant) instances. Their results demonstrated a significant reduction in the screening workload in 11 out of the 15 reviews. Matwin et al. [26] explored the use of a