Maintaining Curated Document Databases Using a Learning to Rank Model: The ORRCA Experience

Iqra Muhammad[1], Danushka Bollegala[1], Frans Coenen[1], Carrol Gamble[2], Anna Kearney[2], and Paula Williamson[2]

[1] Department of Computer Science,
The University of Liverpool, Liverpool L69 3BX, UK
[2] Department of Biostatistics, Institute of Translational Medicine,
The University of Liverpool, Liverpool L69 3BX, UK

Abstract. Curated Document Databases (CDDs) play a critical role in helping researchers find relevant articles in available literature. One such database is the ORRCA (Online Resource for Recruitment research in Clinical trials) database. The ORRCA database brings together published work in the field of clinical trials recruitment research into a single searchable collection. Document databases, such as ORRCA, require year-on-year updating as further relevant documents become available on a continuous basis. The updating of curated databases is a labour intensive and time consuming task. Machine learning techniques can help to automate the update process and reduce the workload needed for screening articles for inclusion. This paper presents an automated approach to the updating of CDDs. The proposed automated approach is founded on a learning to rank model. The approach is evaluated using the ORRCA CDD. Data from the pre-2015 ORRCA CDD was used to train the learning to rank model, and data from the ORRCA 2015 and 2017 updates was used to evaluate the performance of the model. The evaluation demonstrated that significant resource savings can be made using the proposed approach.

Keywords: Learning to Rank, Curated Document Databases

1 Introduction

There is an abundance of scientific research published in the form of academic papers; the number of published papers in most domains has increased in recent years. This makes it challenging for researchers to maintain an overview of the published literature and to find relevant documents in their domain of interest. One solution is the use of Curated Document Databases (CDDs) which bring together, into a single scientific literature repository, all published work in a particular domain. One example of such a CDD is the Online Resource for Recruitment research in Clinical trials (ORRCA[1]) database [6]. The ORRCA

[1] https://www.orrca.org.uk/

database brings together abstracts of papers concerned with recruitment strategies for clinical trials. A manual systematic search process was used to create this database and a similar process is required to update the database on an annual basis. The manual search process for updating the ORRCA database requires substantial human resources. There is a growing need to regularly update this repository as the number of articles being published in the clinical trials domain is growing exponentially. A similar challenge is encountered when updating CDDs in the wider context.

This paper presents an approach to support the automated updating of CDDs. The proposed approach is founded on the use of machine learning, particularly the concept of learning to rank models [10, 11]. The idea is to first obtain a collection of candidate documents from an appropriate bibliographic database. In the case of the ORRCA database candidate documents were obtained by searching over various medical literature bibliographic databases. The next step is then to apply a learning to rank mechanism on the set of candidate documents. One such approach was presented in Norman et al. [12] where a pointwise document ranking algorithm was proposed, the CN algorithm. In this paper, we propose to use the CN algorithm as a foundation and incorporate a Support Vector Regression (SVR) [12] mechanism to rank documents (abstracts). For the evaluation presented in this paper the proposed model was trained using the pre-2015 ORRCA database, and tested using ORRCA 2015 and 2017 updates. The evaluation demonstrated that significant resource savings can be made using the proposed approach.

The remainder of this paper is organised as follows. A brief literature review is given Section 2. Then, in Section 3, a review of the proposed approach is presented. Section 4 considers the necessary data pre-processing. The conducted evaluation of the approach is reported on in Section 5. The report is concluded in Section 6.

2 Literature Review

CDDs require regular updating. This updating process involves considerable human resource as it is typically conducted manually in the form of a systematic review of a candidate collection of documents. The resource required for such systematic review can be significantly reduced by pruning the set of candidates using document ranking. The main objective of document ranking, also referred to as "score-and-sort", is to compute a relevance score for each document and then generate an ordered list of documents so that the top k most relevant documents can be selected. Document ranking models can typically be categorised as being either: (i) probabilistic models or (ii) Learning to Rank Models (LETOR).

Probabilistic ranking models focus on document relevance by assigning a probabilistic value to each document [14]. The idea is to estimate the probabilistic relevance of each document with respect to some criteria (such as a query). Okapi BM25 is an example of a probabilistic ranking model. Okapi BM25 is used as a baseline in many document ranking applications and it has been shown to

substantially outperform alternatives [2]. The main advantage of the probabilistic approach is that it is straightforward. A criticism of the approach is that it is not sufficiently nuanced to capture sophisticated relationships between the prescribed criteria and free text documents.

LETOR use machine learning techniques to "learn" a ranking model. They address the criticism directed at probabilistic ranking models, but feature the disadvantage that training data is usually required. The majority of LETOR fall within two groups: (i) representation-based LETOR and (ii) interaction-based LETOR. Representation based LETOR use representations of the criteria and documents that are independent of each other [14]; while the query and document representations are closely related in interaction-based LETOR, which has been shown to be beneficial for relevance matching between specified criteria and the document collection of interest [2]. A number of such methods for updating curated databases in *domain-specific settings*, have been proposed including: Voting Perceptrons, Lambda-Mart, Decision Trees, SVR, Waode, kNN, Rocchia, hypernym relations, Linear Models, Convolutional Neural Networks, Gradient Boosting Machines, Random Indexing and Random Forests [1, 3, 4, 8, 9, 13, 17]. More recent work has been directed at *non domain-specific settings*. Examples include: (i) iterative learning on implicit feedback [15], (ii) pre-trained sequence to sequence models [11], (iii) transformers based re-ranking [10] and (iv) BERT-based document ranking [16]. Each of these examples is discussed in some further detail below.

In [15] a non-domain specific LETOR was presented, founded on the LambdaMART algorithm, which featured an iterative learning and implicit feedback loop. The implicit feedback from users was used to understand the intent of selecting a document. This feedback was then used to iteratively improve the quality of the learning-to-rank model. The proposed LETOR was trained and tested on a community question answering dataset.

There has been significant recent attention directed at pre-trained document representation models. Pre-trained sequence-to-sequence models generated using Deep Transformer Networks have been applied with respect to a number of Natural Language Processing (NLP) applications including non-domain specific document ranking. Because of the computational complexity involved such models are typically used to re-rank document lists. The work in [11] was directed at adapting a pre-trained sequence-to-sequence model for document ranking (as opposed to re-ranking). The authors demonstrated that, given sufficient training data, their proposed approach out-performed more traditional machine learning based approaches

Deep Transformer Networks, as noted above, are computational expensive and require large amounts of training data. In [10] a transformer based re-ranking approach was presented, the Precomputing Transformer Term Representation (PreTTR) approach, designed to reduce query-time latency that is a feature of Deep Transformer Networks. The idea presented was to pre-compute part of the document representation at "indexing time" (prior to consideration of any document query), and then at "query time" to merge the representation for

a given query with the pre-computed representation and use this to obtain a document re-ranking. As a result the system can be used in real-time learning to rank scenarios.

The most frequently referenced pre-trained NLP models are founded on the Bidirectional Encoder Representations from Transformers (BERT) technique. A recent example where BERT has been used to generate a non-domain specific LETOR was presented in [16]. The pre-trained BERT model was tested on two benchmark LETOR datasets: the MS MACRO passage ranking dataset and the TREC 2004 Robust Track dataset. The evaluation results demonstrated how BERT makes use of the surrounding context of words for document ranking. The results using the MARCO and TREC datasets demonstrated that the proposed BERT-based ranking approach produced equivalent results to previous classification based models.

In the context of the biomedical domain, the example domain of interest with respect to this paper, LETORs have been used for biomedical information retrieval. Recent example can be found in [7], [12] and [5]. The work in [7] was directed at the CLEF (Conference and Labs Evalutaion Forum) 2017 e-Health Lab Task 2. The goal was to automatically rank studies, according to title and abstract, instead of conducting systematic reviews. The approach involved the use of random forests trained on a dataset of 15 systematic reviews conducted by the Oregon Evidence based Practice Center (EPC). The reported evaluation indicated a high recall. In [9] word embeddings and logistic regression were used to create a model for the automated screening of citations. The proposed model was tested on four benchmark datasets from medical domains and demonstrated 100% sensitivity on two of the datasets. The screening workload was reduced by 53%. The work in [12] made use of a logistic regression based ranking model applied to a clinical outcomes document dataset[2]. The proposed approach resulted in a workload reduction of at least 75% and also only missed some 2% of references. This last LETOR has been used as the foundation for the work presented in this paper.

3 The Curated Document Database Update Approach

This section presents an overview of the proposed CDD update approach founded on a SVR Learning to Rank Model (LETOR). A schematic of the approach is given in Figure 1. The start point, top left, is an existing CDD such as the ORRCA CDD. Learning to rank algorithms require pre-labelled training data, both positive examples ("relevant" abstracts) and negative examples ("not relevant" abstracts). By definition CDDs do not include negative examples and thus the positive examples within the CDD need to be augmented with negative examples. For the evaluation presented later in this paper, the pre-2015 version of the ORRCA database was used and augmented with negative examples ("not relevant" documents). The negative examples were taken from the original candidate corpus, which was still available, used to first create the ORRCA CDD.

[2] http://www.comet-initiative.org/

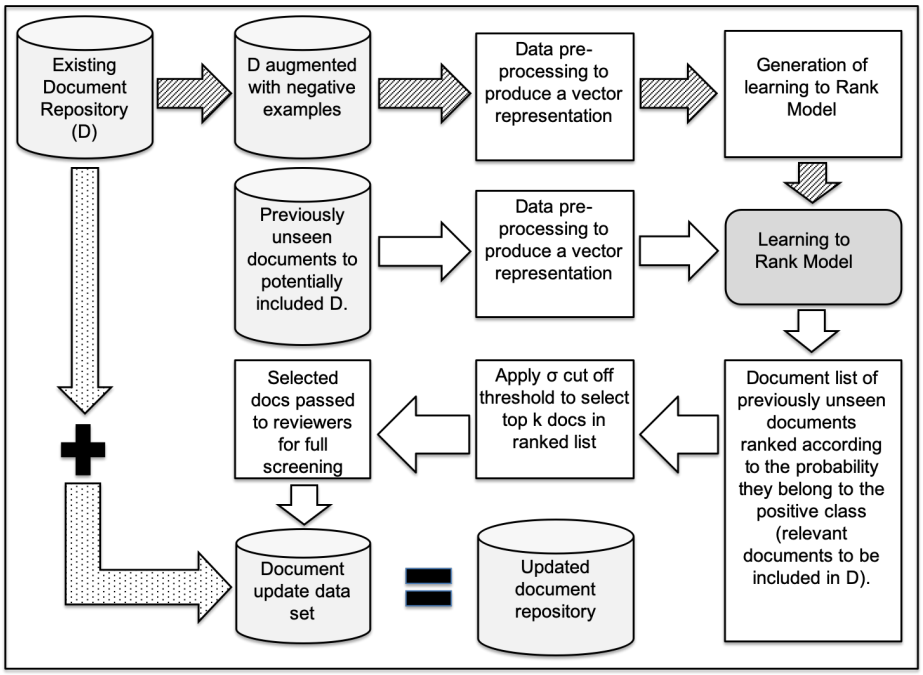

Fig. 1. Schematic of CN Document Collection Update System

This training data, referred to as the pre-2015 Training dataset, was then pre-processed so that it was in a format that allowed for the generation of the desired LETOR; the nature of the pre-processing is discussed in more detail in Section 4 below. A SVR learning approach was used to generate the desired LETOR. More specifically the SVR within the Scikit-learn Python machine learning library[3].

Once the model has been generated it can be applied to previously unseen data; in other words a collection of candidate documents for inclusion in the CDD. For the evaluation presented later in this paper, the 2015 and 2017 OR-RCA update data collection were used (extracted from various medical repositories). SVR is a supervised machine learning model. This model is used to assign a probability value p to a previously unseen document, the probability that the unseen document belongs to the positive class. The probability that the document belongs to the negative class is then $p-1$. A decision boundary, defined in terms of a threshold σ, is then required to assign an appropriate class label c to the previously unseen records. If $p > \sigma$, $c = relevant$; otherwise $c = not\ relevant$ (Equation 1). The challenge is to identify an appropriate value for σ.

[3] Sklearn is a python based machine learning library that contains implementations for various machine learning algorithms.

$$c = \begin{cases} relevant, & if\ p > \sigma \\ not\ relevant, & otherwise \end{cases} \quad (1)$$

For the ORRCA application considered here, the ranking model was trained using log loss[4]. Thus when using the resulting model, each previously unseen record is assigned a probability value between 0 and 1, which in turn can be used to perform a simple binary classification using a decision threshold σ ($0 \leq \sigma \leq 1$). If $\sigma = 0$ is selected all records will be selected as belonging to the positive class; if $\sigma = 1$ is used the likelihood is that no records will be selected (unless we have records where $p = 1$). In the proposed process (Figure 1) it is assumed that human intervention will still be required at the final stage of the process. The value for σ therefore needs to be selected so that the likelihood of false positives is minimised to limit the resource required for the human intervention (the assumption is that the data set selected for further human screening, the screening data set, will always contain some false positives).

4 Pre-processing and Feature Exatraction

Any dataset used for machine learning has to be pre-processed so that the presence of any anomalous records is addressed. In the context of the ORRCA data the pre-processing involved: (i) punctuation and stop word removal and (ii) removal of duplicate records (duplicated abstracts with the same title and content).

The next stage was to transform the data into a format appropriate to the selected learning algorithm to be applied. Usually this is in the form of a feature vector representation. There are various ways that features can be extracted from document collections. With respect to the variation of the CN algorithm considered here the feature extraction involved the extraction of n-grams from the titles and abstracts of documents. An n-gram is defined as contiguous sequence of n words from a given sample of text. Three kinds of n-grams were extracted as potential features: unigrams, bigrams and trigrams. For each n-gram found in a document the Term-Frequency - Inverse Document Frequency (TF-IDF) value was calculated. The TF-IDF value is an indicator of how important a n-gram is to a particular document in a collection or corpus of documents. The TF-IDF value for a n-gram w, $tfidf(w)$, is calculated as shown in Equations 2, 3 and 4 where: (i) tf is term frequency, (ii) w is a given n-gram (iii) d is a document, (iv) $|d|$ is the size of the document d in terms of number of words, (v) D is the entire document collection ($d \in D$) and (vi) $|D|$ is the size of D in terms of number of documents. Once the TF-IDF values have been calculated a threshold θ is required to decided which n-grams go into the vector representation. For the evaluation presented in Section 5, the default value for θ in the Scikit-learn

[4] Logarithmic loss (related to cross-entropy) measures the performance of a classification model where the prediction output is a probability value between 0 and 1

implementation was used. The ORRCA pre-2015 Training dataset was unbalanced hence, in order to mitigate against the imbalance between the number of positive and negative examples, the training weight for the positive examples was increased to 80.

$$tfidf(w) = tf(w) \times idf(w) \tag{2}$$

$$tf(w) = \frac{frequency\ count\ of\ w \in d}{|d|} \tag{3}$$

$$idf(w) = \frac{|D|}{total\ number\ of\ d \in D\ in\ which\ w\ appears} \tag{4}$$

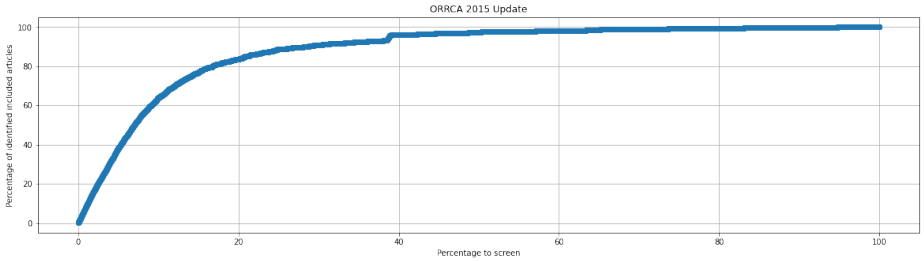

Fig. 2. Effort-gain curve for ORRCA 2015 dataset with both titles and abstracts, showing the percentage of articles screened versus the percentage of relevant articles identified.

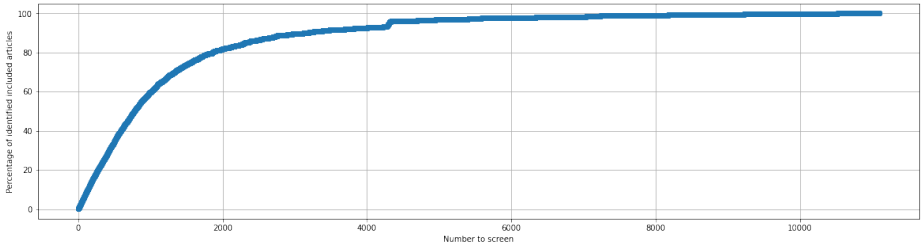

Fig. 3. Effort-gain curve for ORRCA 2015 dataset with both titles and abstracts, showing the number of articles screened versus the percentage of relevant articles identified.

5 Evaluation

In this section an analysis of the proposed LETOR-based CDD update approach is presented. For the analysis, as already noted above, the ORRCA pre-2015

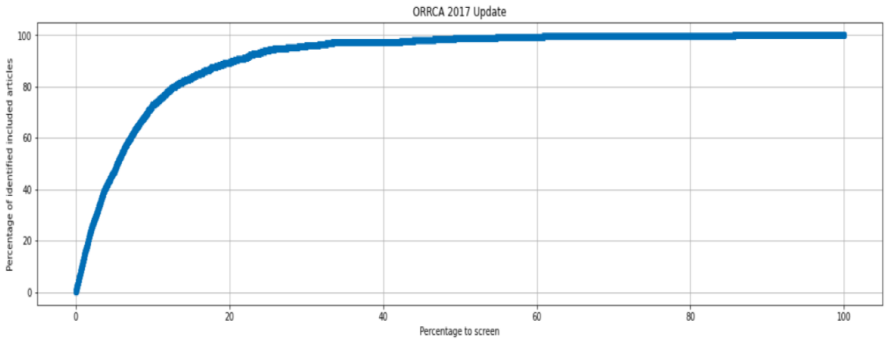

Fig. 4. Effort-gain curve for ORRCA 2017 dataset with both titles and abstracts, showing the percentage of articles screened versus the percentage of relevant articles identified.

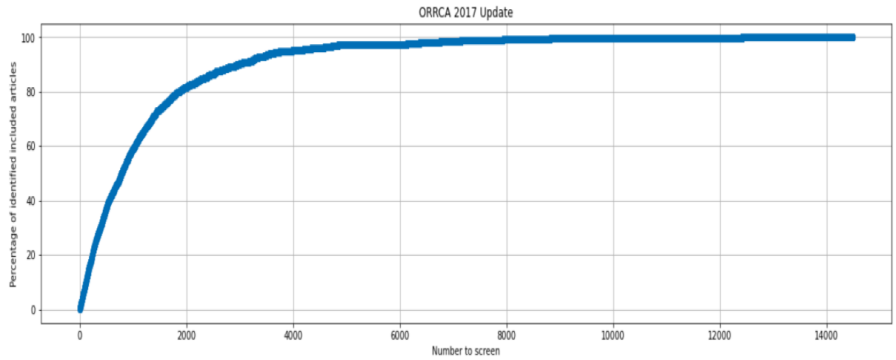

Fig. 5. Effort-gain curve for ORRCA 2017 dataset with both titles and abstracts, showing the number of articles screened versus the percentage of relevant articles identified.

database, augmented with negative examples, was used as the training set and the 2015 and 2017 ORRCA updates as the test set. Some statistics concerning these two data set are given in Table 1. The objectives of the analysis were to:

1. Determine a most appropriate value for σ.
2. Measure the performance of the proposed CDD update approach with respect to: (i) effectiveness and (ii) time saving.

With respect to the first objective, effort-gain curves were used where the number of documents, or the percentage of documents, in the ranked list of documents selected for manual screening (the *effort*) was plotted against and number/percentage of true-positives identified (the *gain*). These curves tend to be exponential in nature, they feature an initial steep climb and then a flattening-off.

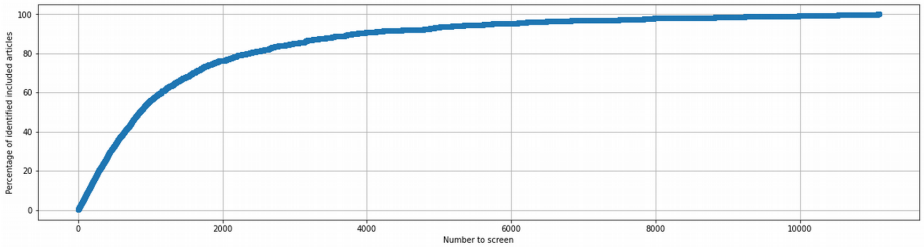

Fig. 6. Effort-gain curve for ORRCA 2015 dataset with titles only, showing the number of articles screened versus the percentage of relevant articles identified.

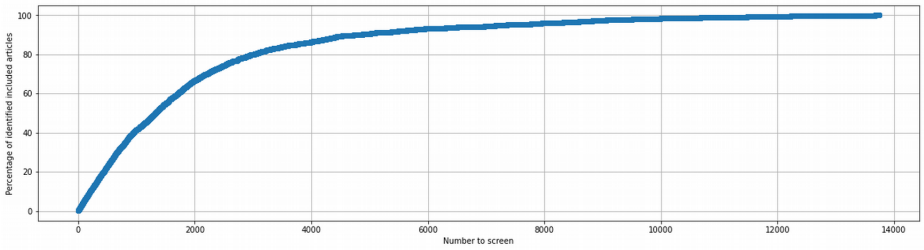

Fig. 7. Effort-gain curve for ORRCA 2017 dataset with titles only, showing the number of articles screened versus the percentage of relevant articles identified.

The idea is that σ should be selected according to the location of the elbow between the climb and the flattening-off because this is the point where the "gain" starts to decrease compared to the "effort" required for manual screening. With respect to the second objective, the metrics used to measure effectiveness were precision and recall, calculated as given in Equations 5 and 6 where: (i) TP is the number of true positives, (ii) FP is the number of false positives and (iii) FN us the number of false negatives. A true positive is an outcome where the model correctly predicts the positive class. A true negative is an outcome where the model correctly predicts the negative class. A false negative is an outcome where the model incorrectly predicts the negative class. Recall that we wish to select a value for σ whereby the set of selected abstracts to be screened by a human is comprised of as many relevant abstracts (true positives) as possible.

Database Name	Positive Examples		Negative Examples		Total
	Num.	%	Num.	%	
Pre-2015 Training dataset	4570	8.2	51460	91.8	56030
ORRCA 2015 Update Test dataset	1302	11.7	9797	88.3	11099
ORRCA 2017 Update Test dataset	1027	7.1	13458	92.9	14485

Table 1. Statistical overview of the ORRCA training and test data

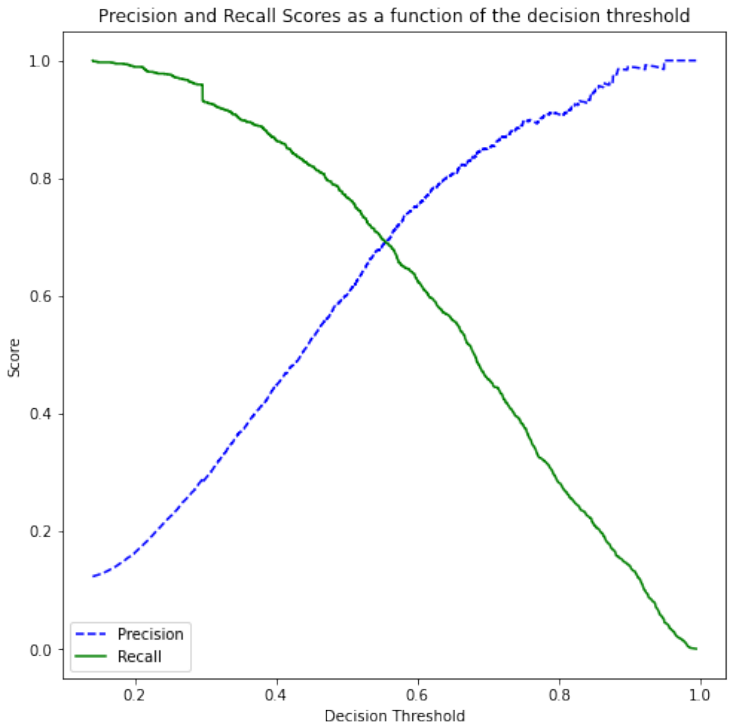

Fig. 8. Precision-recall curve for ORRCA 2015 dataset with both titles and abstracts, showing the decision thresholds σ values on the x-axes and the precision and recall values on y-axes

$$Precision = TP/(TP + FP) \tag{5}$$

$$Recall = TP/(TP + FN) \tag{6}$$

The resulting effort-gain curves are presented in Figures 2, 3, 4 and 5. Figures 2 and 3 show the curves for the 2015 update, and Figures 4 and 5 the curves for the 2017 update, based on features from titles and abstracts of documents. For the effort-gain curves given in Figures 2 and 4 the x-axis denote the percentage of candidate abstracts (thus both relevant and irrelevant abstracts) to be screened as a percentage of the total number of abstracts; whilst Figures 3 and 5 show the same information but in terms of absolute numbers. Figure 6 and 7 shows the effort-gain curves for the 2015 update and 2017 update based on features from titles of documents only.

Inspection of Figure 2, titles and abstract of documents for the 2015 update, indicates that 97% of relevant abstracts can be identified by considering the top 45% of abstracts in the ranked document list. From Figure 3 this equates to the top 4500 abstracts in the ranked document list. Similarly, inspection of

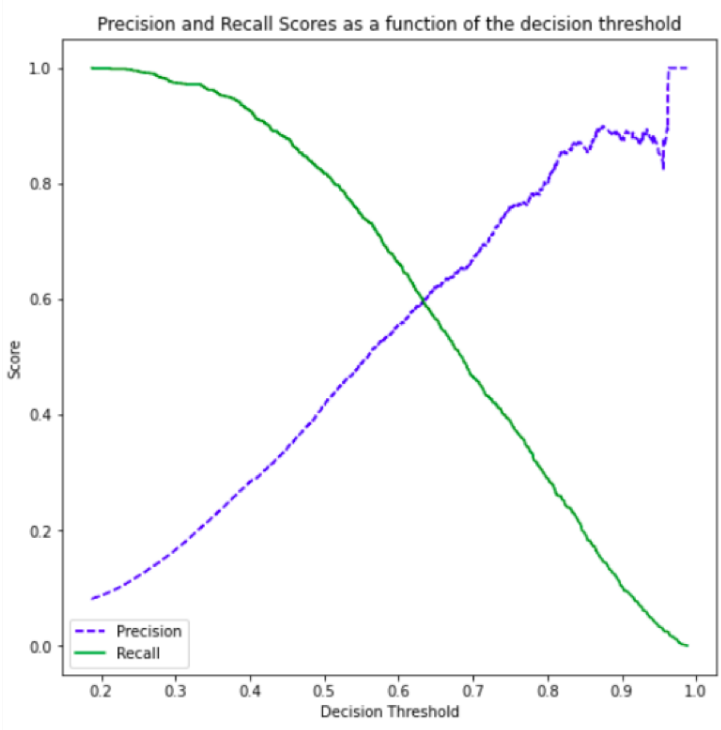

Fig. 9. Precision-recall curve for ORRCA 2017 dataset with both titles and abstracts, showing the decision thresholds σ values on x-axes and precision and recall values on the y-axes

Figure 4, titles and abstract of documents for the 2017 update, indicates that identification of 97% of relevant abstracts requires consideration of the top 40% of abstracts in the ranked document list; and that, from Figure 5, this equates to the top 6000 abstracts. Thus screening the top 40%-45% of abstracts in the ranked abstracts list would result in a loss of roughly 3% of the relevant abstracts; which, it is argued here, is an acceptable compromise (trade-off) between gain and effort. Selection of the top 40%-45% abstracts, in this case, would equate to a σ threshold of $\sigma = 0.3$.

Inspection of Figure 6, titles only for the 2015 update, indicates that 97% of relevant abstracts can be identified by considering the top 6000 abstracts in the ranked document list. Similarly, inspection of Figure 7, titles only for the 2017 update, indicates that 97% of relevant abstracts can be identified by considering the top 8000 abstracts in the ranked document list. This indicates that screening with titles of documents only is not a feasible option for identification of relevant documents in a ranked list of documents in an acceptable time frame, whereas screening with both titles and abstracts is a more practical option for the automatic screening of documents.

In order to estimate the time saved by automating the manual screening process, the assumption was made that the screening rate of an experienced screener is one abstract per minute. Using $\sigma = 0.3$, and considering the 2015 ORRCA update (titles and abstracts), this will result in 5099 ($11099 - 6000 = 5099$) abstracts being excluded, equating to a time saving of $5099 \div 60 = 85.0$ hours (assuming an experienced screener for the abstract screening process). With respect to the 2017 ORRCA update (titles and abstracts), by selecting $\sigma = 0.3$ 8485 abstracts would be excluded ($14485 - 6000 = 8485$) equating to a time saving of $8485 \div 60 = 141.4$ hours.

Figures 8 and 9 show the precision-recall curve for the ORRCA 2015 and 2017 updates (titles and abstracts) respectively. In these figures σ is plotted on the x-axis and the precision/recall score on the y-axis. Each figure features two curves, one for precision and one for recall. As noted earlier, there is a trade off between the number of relevant abstracts and the number of irrelevant abstracts in the data set identified for screening depending on the selected value for σ. This is illustrated in the figure. From Figures 8 and 9, in order to achieve a recall of 1 (identification of all relevant abstracts) $\sigma = 0.2$ would be required. Thus, after screening the first 40%-45% of the candidate abstracts, equivalent to a $\sigma = 0.3$ and identifying 97% of the relevant abstracts, we would be prepared to accept a loss of 3%; this is argued to be an acceptable trade-off in an automated systematic review system.

6 Conclusion

In this paper, a Curated Document Database (CDD) update approach has been presented founded on the concept of learning to rank. More specifically using a Support Vector Regression (SVR) Learning to Rank Model (LETOR). The proposed approach was applied to the task of updating the Online Resource for Recruitment research in Clinical trials (ORRCA) database, but is equally applicable to alternative CDDs. For the evaluation the ORRCA pre-2015 database, augmented with negative examples, was used as the training set, and the 2015 and 2017 ORRCA update candidate dataset as the test set. The evaluation results obtained demonstrate that, when considering both titles and abstracts and using $\sigma = 0.3$, 97% of relevant abstracts can be identified by considering the top 40%-45% of potential abstracts ranked according to probability that they belong to the positive class. This equated to a time saving of some 85 to 141 hours. Experiments using titles only indicated that this was not an appropriate approach and that to obtain adequate recall both titles and abstracts needed to be considered. More generally, the results indicated that automated screening can be used to reduce the workload associated with CDD updating. In future, the authors intend to investigate the use of entity embeddings and BERT based learning-to-rank models for CDD updating.

References

1. Elaine Beller, Justin Clark, Guy Tsafnat, Clive Adams, Heinz Diehl, Hans Lund, Mourad Ouzzani, Kristina Thayer, James Thomas, Tari Turner, et al. Making progress with the automation of systematic reviews: principles of the international collaboration for the automation of systematic reviews (icasr). *Systematic reviews*, 7(1):1–7, 2018.
2. Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74, 2017.
3. Karen G Dowell, Monica S McAndrews-Hill, David P Hill, Harold J Drabkin, and Judith A Blake. Integrating text mining into the mgi biocuration workflow. *Database*, 2009, 2009.
4. Brian E Howard, Jason Phillips, Kyle Miller, Arpit Tandon, Deepak Mav, Mihir R Shah, Stephanie Holmgren, Katherine E Pelch, Vickie Walker, Andrew A Rooney, et al. Swift-review: a text-mining workbench for systematic review. *Systematic reviews*, 5(1):87, 2016.
5. Evangelos Kanoulas, Dan Li, Leif Azzopardi, and Rene Spijker. Clef 2017 technologically assisted reviews in empirical medicine overview. In *CEUR Workshop Proceedings*, volume 1866, pages 1–29, 2017.
6. Anna Kearney, Nicola L Harman, Anna Rosala-Hallas, Claire Beecher, Jane M Blazeby, Peter Bower, Mike Clarke, William Cragg, Sinead Duane, Heidi Gardner, et al. Development of an online resource for recruitment research in clinical trials to organise and map current literature. *Clinical Trials*, 15(6):533–542, 2018.
7. Madian Khabsa, Ahmed Elmagarmid, Ihab Ilyas, Hossam Hammady, and Mourad Ouzzani. Learning to identify relevant studies for systematic reviews using random forest and external information. *Machine Learning*, 102(3):465–482, 2016.
8. Martin Krallinger, Miguel Vazquez, Florian Leitner, David Salgado, Andrew Chatr-Aryamontri, Andrew Winter, Livia Perfetto, Leonardo Briganti, Luana Licata, Marta Iannuccelli, et al. The protein-protein interaction tasks of biocreative iii: classification/ranking of articles and linking bio-ontology concepts to full text. *BMC bioinformatics*, 12(S8):S3, 2011.
9. Ivan Lerner, Perrine Créquit, Philippe Ravaud, and Ignacio Atal. Automatic screening using word embeddings achieved high sensitivity and workload reduction for updating living network meta-analyses. *Journal of clinical epidemiology*, 108:86–94, 2019.
10. Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. Efficient document re-ranking for transformers by precomputing term representations. *arXiv preprint arXiv:2004.14255*, 2020.
11. Rodrigo Nogueira, Zhiying Jiang, and Jimmy Lin. Document ranking with a pre-trained sequence-to-sequence model. *arXiv preprint arXiv:2003.06713*, 2020.
12. Christopher R Norman, Elizabeth Gargon, Mariska MG Leeflang, Aurélie Névéol, and Paula R Williamson. Evaluation of an automatic article selection method for timelier updates of the comet core outcome set database. *Database*, 2019, 2019.
13. Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):5, 2015.
14. Gaurav Pandey. Utilization of efficient features, vectors and machine learning for ranking techniques. *JYU dissertations*, 2019.

15. Mateus Pereira, Elham Etemad, and Fernando Paulovich. Iterative learning to rank from explicit relevance feedback. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 698–705, 2020.
16. Yifan Qiao, Chenyan Xiong, Zhenghao Liu, and Zhiyuan Liu. Understanding the behaviors of bert in ranking. *arXiv preprint arXiv:1904.07531*, 2019.
17. Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, et al. Overview of the clef ehealth evaluation lab 2018. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 286–301. Springer, 2018.