Document Ranking for Curated Document Databases using BERT and Knowledge Graph Embeddings: Introducing GRAB-Rank

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 (CDD) play an important role in helping researchers find relevant articles in scientific literature. Considerable recent attention has been given to the use of various document ranking algorithms to support the maintenance of CDDs. The typical approach is to represent the update document collection using a form of word embedding and to input this into a ranking model; the resulting document rankings can then be used to decide which documents should be added to the CDD and which should be rejected. The hypothesis considered in this paper is that a better ranking model can be produced if a hybrid embedding is used. To this end the Knowledge Graph And BERT Ranking (GRAB-Rank) approach is presented. The Online Resource for Recruitment research in Clinical trials (ORRCA) CDD was used as a focus for the work and as a means of evaluating the proposed technique. The GRAB-Rank approach is fully described and evaluated in the context of learning to rank for the purpose of maintaining CDDs. The evaluation indicates that the hypothesis is correct, hybrid embedding outperforms individual embeddings used in isolation. The evaluation also indicates that GRAB-Rank outperforms a traditional approach based on BM25 and and a ngram-based SVR document ranking approach.

Keywords: BERT · knowledge graph concepts · Document ranking.

1 Introduction

The number of published papers in scientific research is increasing rapidly in any given domain. Consequently, researchers find it difficult to keep up with the exponential growth of the scientific literature. In order to address this challenge many organisations manage Curated Document Databases (CDDs). CDDs are specialised document collections that bring together published work, in a defined domain, into a single scientific literature repository. One example of such a CDD, and that used for illustrative purposes in this paper, is the Online Resource for

Recruitment research in Clinical trials (ORRCA[1]) CDD [7]. The ORRCA CDD brings together abstracts of papers concerned with the highly specialised domain of recruitment strategies for clinical trials.

The provision of CDDs provide a useful facility for researchers. However, for CDDs to remain useful, they must be constantly updated, otherwise their utility is of only temporary value. The challenge is illustrated in the context of the ORRCA CDD in Figure 1. From the figure the exponential, year-on-year, growth of the number of papers can be observed clearly. Updates are conducted using what is referred to as a *systematic review process*. The systematic review is typically conducted manually by querying larger document collections, a time consuming task. In the case of ORRCA, the PubMed search engine for the MEDLINE life sciences and biomedical abstracts database was used for the systematic review. The process can be enhanced using *document ranking* so that candidates for an update can be ranked according to relevance and the top k considered in more detail, whilst the remainder can be rejected.

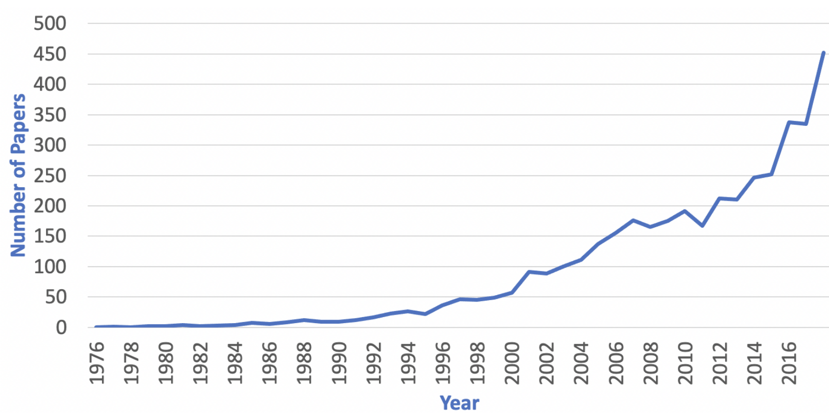

Fig. 1. ORRCA papers and articles 1976-2017, illustrating the exponential growth of the number of publications directed at recruitment strategies for clinical trials.

Document ranking has been extensively used in the context of document retrieval. The traditional approach, given a particular search query, is to rank documents using a frequency measure that counts the frequency whereby terms in the search query appear in each candidate document [6, 18]. However, frequency based document ranking models fail to capture the semantic context behind individual search queries. An alternative is to use a learning to rank model more suited to capturing the semantic meaning underpinning search queries [14, 16]. Recent work on Learning to Rank (LETOR) has used word embeddings of various kind as the input [11, 14]. Word embeddings can be learnt from scratch or a pre-trained embedding model can be adopted. Existing word embedding based approaches to LETOR have focused on a single embedding, with good results; a popular choice is to use Bidirectional Encoder Representations from

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

Transformers (BERT) embeddings [10, 17]. The intuition presented in this paper is that a *hybrid approach* using two orthogonal, but compatible, embeddings will result in a more effective ranking (in the context of the CDD update problem). An intuition that is supported by the observations given in [1]. To this end this paper presents the Knowledge Graph And BERT Ranking (GRAB-Rank) approach to LETOR, designed to support the periodic updating of CDDs, that combines BERT word embeddings and knowledge graph concept embeddings; the latter generated using a bespoke *random walk* technique.

The GRAB-Rank approach is fully described and evaluated. The proposed approach assumes the availability of a literature knowledge graph. Techniques whereby document knowledge graphs can be constructed, given a document corpus, are available (see for example [13]). The presented evaluation was conducted using the ORRCA CDD. GRAB-Rank results were compared with an approach based on the popular Okapi BM25 ranking function [23] and earlier work directed at the updating of the ORRCA CDD as reported in [16] where a Support Vector Regression (SVR) based technique was presented. It was found that GRAB-Rank produced better results than when either of the considered embeddings were used in isolation, and that the proposed hybrid embedding model outperformed the BM25 and ngram-based SVR document ranking comparator approaches.

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. In this paper a mechanism for updating the ORRCA CDD [7] is presented founded on a hybrid document ranking technique. Recent work in document ranking has been focused on using external knowledge for improving document rankings. Especially the use of contextualised models such as BERT. LETOR models can be categorised as being either: (i) Traditional document ranking models, (ii) Semantic document ranking models, or (iii) Knowledge graph document ranking models. Sub-sections 2.1, 2.2 and 2.3 give further detail with respect to each of these categories.

2.1 Traditional Document Ranking Models

Traditional document ranking models are founded on statistical or probabilistic approaches. Many variants of these methodologies have been proposed and continue to be proposed. Most are founded on a vector space model of the input document corpus where the dimensions of the vector space are defined using terms that appear in the document collection. The terms to be included are typically selected using a scoring mechanism. Term Frequency - Inverse Document Frequency (TF-IDF) is a popular choice [11]. In this manner a n-dimensional vector space can be constructed. A popular algorithm for generating vector representations of words is GloVE (Global Vectors for Word Representation), an unsupervised learning algorithm that operates by aggregating global word-word co-occurrence statistics found in an input corpus [20]. An alternative mechanism of generating a vector space model is to use *word n-grams*. This was the technique used in [12] and [16]. The significance of the techniques used in [12] and [16] is that it was evaluated using the ORRCA CDD, and hence used with respect to the evaluation presented later in this paper to compare with the operation of GRAB-Rank. Once the input document corpus has been converted to a vector based representation, a document ranking function can be applied to rank the documents in decreasing order of relevance to a query. A popular ranking function is the Okapi BM25 ranking function which is founded on a probabilistic retrieval framework [23]. The Okapi BM25 function was also adopted as a document ranking baseline with respect to the work presented in this paper.

2.2 Semantic Document Ranking Models

The traditional statistical and probabilistic document ranking models assume each term is independent of its neighbours. Semantic document ranking models take into account the context of terms in relation to their neighbouring terms, in other words the "semantic" context associated with each term. We refer to this using the phrase *word embedding*. The distinction can be illustrated by considering the word "bank"; using a semantic context representation this would comprise a number of vectors depending on the context of the word "bank", either as: (i) an organisation for investing and borrowing money, (ii) the side of a river or lake, (iii) a long heap of some substance, (iv) the process of heaping up some substance or (v) the process of causing a vehicle to tilt to negotiate a corner. Using a non-contextualised representation the word "bank" would be represented using a single vector regardless of context.

Semantic representations are generated using a contextual model to generate the desired word embeddings; different terms that have the same semantic meaning are thus represented in a similar way. The required contextual model can be learned directly, typically using deep learning, from the document corpus of interest. Examples of document ranking systems that use a learnt contextual model to produce a word embedding can be found in [3, 15, 28]. However, learning a contextual model requires considerable resource. The alternative is to use an existing pre-trained contextual model to generate a word embedding for a given corpus. A popular choice of pre-trained contextual model is the Bidirectional Encoder Representations from Transformer (BERT). BERT takes into account the context of a target word using the surrounding words in a large corpora; BERT has been used with respect to many downstream natural language processing tasks including document ranking [15, 26]. An alternative contextual model that can be used is the embeddings from Language Model ELMo [21].

This model is based on deeply contextualized word embeddings which are created from Language Models (LMs). BERT is a transformer-based architecture while ELMo is Bi-LSTM Language model. BERT is purely Bi-directional and ELMo is semi-bidirectional. However, with respect to the work presented in this paper, because of BERT's popularity and its ease of use in Python, a BERT pre-trained model based sentence embeddings were used for the downstream task of ranking scientific abstracts.

2.3 Knowledge Graph Document Ranking Models

The work presented in this paper assumes CDDs represented as literature knowledge graphs. A knowledge graph is a collection of vertices and edges where the vertices represent entities or concepts, and the edges represent a relationship between entities and/or concepts. The reason for using knowledge graphs is that they provide efficient storage and retrieval in the context of linked descriptions of data. Some well known examples of knowledge graphs include Freebase [2] and YAGO [22]. In the context of document knowledge graphs the concepts stored at vertices represent semantic information which, it is argued here, can be used in the form of knowledge graph embeddings for document ranking purposes. Examples of recent work directed at knowledge graphs for document ranking include the entity-based language models described in [8, 9, 27]. This existing work has demonstrated the viability of knowledge graph based document ranking. The work presented in this paper proposes a hybrid approach that combines semantic document ranking with knowledge graph document ranking.

3 Problem Definition

A CDD is a data set of the form $D = \{d_1, d_2, \dots d_n\}$ where each $d_i \in D$ is a document (research article/paper). For the CDD to remain useful it is necessary for it to be periodically updated by adding the set of recently published new documents Q to D so that $D_{new} = D \cup Q$. The set Q is traditionally generated using a systematic review process [7] applied to a larger data set U ($Q \subset U$). Systematic reviews involve a detailed plan and search strategy with the objective of identifying, appraising, and synthesizing all relevant studies on a particular topic [7, 25] . The challenge is to automatically generate Q in such a way that Q contains as many relevant documents as possible. The anticipation is that it will not be possible to automatically generate a set Q that contains all relevant documents and no irrelevant documents. The idea is therefore to apply a Learning to rank model (LETOR) whereby U is ordered according to relevance score and the top k documents selected for potential inclusion in D.

4 BERT and knowledge graph embeddings based document ranking

This section presents the proposed Knowledge Graph and BERT Ranking (GRAB-Rank) approach. A schematic of the approach is presented in Figure 2. The input is a collection of documents U to be potentially included in D. The next stage is to generate two sets of document embeddings: (i) document embeddings generated from a random walk of a knowledge graph G generated from U, and (ii) document embeddings generated using BERT. The first requires the transformation of U into G, how this can be achieved is presented in Sub-section 4.1. The process of generating document embeddings from G is then described in Sub-section 4.2. The process for generating document embeddings using BERT is described in Sub-section 4.3. Once we have the two types of document embeddings, these are combined into a single embedding, by concatenating one to the other. The concatenated embedding is then used as input to a LETOR model. With respect to the evaluation presented later in this paper, and as indicated in the figure, a Support Vector Regression (SVR) model was used to generate the document ranking. A SVR model was used because this has been shown to produce good results as evidenced in [12] and [16], a previously proposed approaches for updating CDDs which also focused on the ORRCA CDD. SVR uses the same principle as Support Vector Machines (SVMs) but with respect to regression problems. The SVR LETOR model, once learnt, can be used to assign a ranking value to each document in U. To obtain Q from U we then need a cut-off threshold value σ. The work in [12] reported the results from a sequence of experiments to establish the most appropriate value for σ. They found that 97% of relevant abstracts can be identified by considering the top 40-45% of potential abstracts. This was found to equate to a value of σ=0.30. For the evaluation presented in this paper, σ=0.25 was used (so as to include a "safety margin").

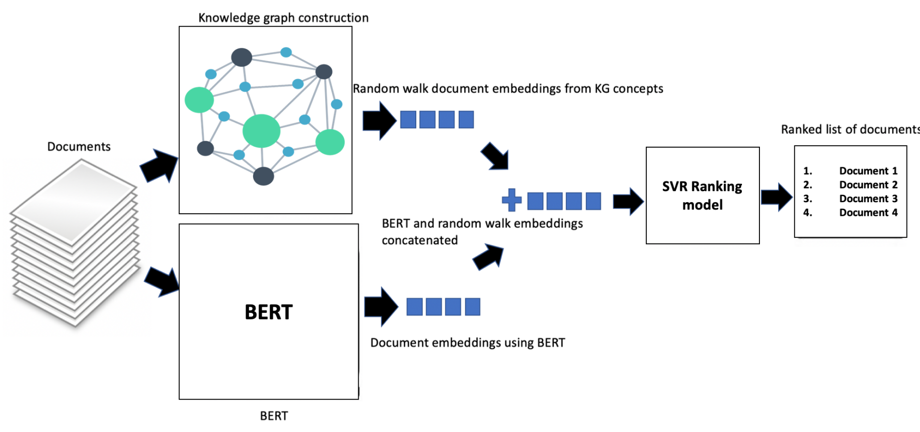

Fig. 2. Schematic of the GRAB-Rank approach.

4.1 Knowledge graph construction

The prerequisite of the GRAB-Rank approach for the maintenance of CDDs is a literature knowledge graph $G = \{V, E\}$ where the set of vertices V represent documents and concepts, and the set of edges E represent relationships between the vertices. There are various mechanisms whereby G can be constructed; it can be done manually, but is clearly better addressed in an automated manner. One proposed solution, and that adopted with respect to this paper, is the OIE4KGC (Open Information Extraction for Knowledge Graph Construction) approach presented in [13]. The OIE4KGC solution commences by extracting concept-relation-concept triples from a given document collection D (a CDD) using the RnnOIE Open Information Extraction (OIE) tool [24]. The triples are then filtered so that only the most relevant concepts are retained, each identified by a unique label. The retained triples, are then used to construct G such that the set of vertices V represents concepts and documents (in the following the terms concept vertex and document vertex are used to distinguish between the two), and the set of edges E comprises either: (i) the extracted relations from one concept to another concept, or (ii) "mention" relationships from a document to a concept. A mention relationship between a document and a concept implies that a document "mentions" this particular concept. Similar concepts in the knowledge graph were linked, using a biomedical entity linker.

4.2 Knowledge Graph Concept Embeddings using a Random walk

This section presents the process for generating concept embeddings from a literature knowledge graph (generated as described above); the first form of embedding in the proposed hybrid embedding approach. The process commences by generating a sequence of random walks (paths) linking concept vertices. The *random walk* idea was first proposed in [19], where it was defined as a sequence of elements creating a path in a mathematical space. Conceptually, a random walk across a graph can be considered as a sequence of vertices. In the case of the proposed approach each vertex in the sentence will be a "concept". Therefore, each walk generated from a concept vertex in G can be interpreted as a natural language sentence comprised of the concepts covered by the walk. The "sentences" can then be processed using a range of text machine learning models, such as the "bag of words" model or the "skip gram" model [5]. Random walks were generated for every concept node in G. It takes a high amount of computational resources to generate random walks for each vertex, hence a number of 100 random walks was chosen for each vertex. The length of each generated walk was restricted to k vertices. For the evaluation reported on in the following section, Section 4, experiments were conducted using a range of values for k, from $k = 1$ to $k = 5$ incrementing in steps of 1.

The foregoing was implemented using the node2vec framework [4] and the skip-gram model. Using the node2vec framework random walks can be generated using a number of strategies, these can be broadly categorised as Breadth-First Sampling (BFS) or Depth-First Sampling (DFS). The breadth-first strategy involves identifying all the immediate neighbours of a current vertex $v_i \in V$, to be included in the random walks to be generated and then moving on to immediate neighbours plus one, and so on until we reach random walks of length k. The depth-first strategy involves generating each entire random walk in turn rather than “in parallel”. A breadth-first strategy (BFS) was used for the proposed GRAB-Rank approach.

4.3 BERT contextualised embeddings

This section discusses the contextualised embedding process, the second embedding used with respect to the proposed GRAB-Rank approach. The idea is to use transfer learning; the process of using a pre-trained deep language model to generate document embeddings. There are a number of such language models available, examples include ELMo [21] and BERT [15]. With respect to the evaluation presented later in this paper the BERT language model was used. The advantage offered by these models, as noted in Section 2, is that they are context aware; unlike many alternative models, such as GloVe [20], where each word is represented using a single vector regardless of context. Using a pre-trained language model, document embeddings are generated by replacing each word in a given document with the corresponding (BERT) word embedding. All the word vectors in the document are then concatenated to obtain a single document embedding. Contextualised language models consist of multiple stacked layers of representation (and an input layer); the greater (deeper) the number of layers the greater the extent of the context incorporated into a word representation. To generate word embeddings all layers can be used or the top n (most significant) layers. With respect to the evaluation presented later in this paper results are reported using all twelve BERT layers.

5 Evaluation

This section presents the evaluation of the proposed GRAB-Rank approach. For the evaluation, the abstracts for the 2015 and 2017 systematic review updates of the ORRCA CDD were used because: (i) a ground truth was readily available (the abstracts eventually selected for inclusion in ORRCA were known); and (ii) ORRCA had been used in previous LETOR studies, namely those presented in [12] and [16], hence a comparison could be conducted. Two data sets were generated using the 2015 and 2017 abstracts:

ORRCA-400 A small data set which could be manually inspected and analysed in the context of the proposed GRAB-Rank approach, referred to as the ORRCA-400 data set because it comprised 400 abstracts, 200 abstracts included in ORRCA and 200 excluded. Thus an even distribution.

ORRCA-Update A much larger data set to test the scalability of the proposed approach made up of the entire 2015 and 2017 ORRCA update collections, $11,099$ abstracts for the 2015 update (1302 included and 9797 excluded) and 14,485 for the 2017 update (1027 included and 13458 excluded).

Both datasets were pre-processed by removing punctuation and stop words. For stop word removal ntlk[2] was used. For training and testing a 60:40 training-testing split was used with respect to both data sets. Approaches similar to

[2] https://www.nltk.org/

GRAB-Rank [10, 15] have used similar splits for training and testing document ranking models.

The objectives of the evaluation were:

1. To conduct an ablation study to compare the operation of the proposed GRAB-Rank approach with using only BERT embeddings and only knowledge graph embeddings, so as to demonstrate that the proposed hybrid approach outperformed the component approaches when used in isolation.
2. To compare the operation of the proposed GRAB-Rank approach with alternative traditional document ranking systems applied directly to the input data, namely: (i) the Okapi BM25 ranking function based approach described in [23]; and (ii) the n-grams approach, with a SVR ranking model, presented in [12] and [16]. Both were discussed previously in Sub-section 2.1.
3. To investigate the effect of the parameter k, the random walk length, on the operation of the GRAB-Rank approach.

The evaluation was conducted using a NVidia K80 GPUs kaggle kernel. The evaluation metrics used were precision and recall. Our dataset was labelled in a binary manner (relevant and not relevant) hence traditional document ranking metrics which require a "ground truth" ranking, such as MAP, MRR and NDCG, could not be used. For generating BERT embeddings all BERT Layers, as suggested in [21], were used. This was because using all layers produces a richer result, at the expense of increased run time. However, runtime is not an issue in the context of CDD maintenance as it is an activity not conducted frequently. In the case of the ORRCA CDD this is typically updated once every two years (because of the significant human resource involved). For the initial experiments conducted with respect to Objectives 1 and 2 above, $\sigma = 0.25$ was used with respect to the SVR LETOR models generated for reasons given previously in Section 5.

The results with respect to the first two objectives are given in Table 1. From the table it can be seen that in all cases the proposed GRAB-Rank hybrid approach produced the best performance with respect to both evaluation data sets and with respect to both recall and precision. From Table 1 it can also be seen that BERT only embedding tended to outperform knowledge graph embedding. The precision values are relatively low for the ORRCA-Update dataset. It is conjectured that this is because the ORRCA-Update dataset (25,584 documents) was significantly larger the ORRCA-400 dataset (400 documents).

To determine the most appropriate value for k experiments were conducted using a range of values for k from $k = 1$ to $k = 5$ incrementing in steps of 1. The results are presented in Table 2. From the table it can be seen that the best precision was obtained when $k = 2$ for the ORRCA-400 dataset, and $k = 3$ for the ORRCA-Update dataset. There was no clear best value for k with respect to recall. From the table it can also be seen that a low value of k ($k < 2$) produced poor recall. This could be attributed to the fact that the higher the value for k the more similar concepts that are included in the knowledge graph embedding and hence the better the recall (greater number of relevant documents at the top of a ranked document list).

Table 1. The performance of GRAB-Rank in comparison with using BERT embeddings or knowledge graph embeddings in isolation, and with using the BM25 and SVR ranking models with n-grams (best results in bold font).

Document ranking technique	ORRCA-400		ORRCA-Update	
	Precision	Recall	Precision	Recall
GRAB-Rank with SVR	**0.81**	**0.50**	**0.26**	**0.88**
BERT embeddings only with SVR	0.76	0.47	0.23	0.80
Knowledge graph embeddings only with SVR	0.75	0.46	0.26	0.87
word2vec vectors with BM25 ranking	0.53	0.33	0.16	0.54
word2vec for n-grams with SVR	0.79	0.49	0.07	0.49

Table 2. The performance of GRAB-Rank with using a range of values for k, the random walk length (best results in bold font).

k	ORRCA-400		ORRCA-Update	
	Precision	Recall	Precision	Recall
1	0.68	0.42	0.17	0.59
2	**0.75**	**0.46**	0.24	0.83
3	0.74	**0.46**	**0.26**	**0.87**
4	0.73	0.45	**0.26**	0.86
5	0.74	**0.46**	**0.26**	0.86

6 Future work and Conclusion

This paper has presented the GRAB-Rank approach to partially automate the process of maintaining CDDs which would otherwise need to be maintained using a manual systematic review process. GRAB-Rank is a LETOR mechanism founded on a hybrid representation comprised of a literature knowledge graph embedding generated using a random walk of length k and a BERT contextual embedding. The hybrid embedding was then used as an input into a LETOR mechanism. For the presented evaluation SVR was used to generate the desired LETOR model. The hypothesis was that a hybrid document embedding approach would produce a better ranking than if the component embeddings were used in isolation. The GRAB-Rank approach was evaluated using two datasets extracted from the data used for the maintenance of the ORRCA CDD. The evaluation results obtained indicated that the hypothesis was correct, hybrid embedding outperforms individual embeddings used in isolation. The operation of GRAB-Rank was also compared with two forms of traditional approach, one based on BM25 and the other on n-gram based SVR. Grab-Rank was shown to outperform the traditional approaches. For future work the authors plan to improve the document ranking model by conducting further experiments, using the ORRCA CDD, with other pre-trained language model such as GPT-2 and GPT-3 for creating document embeddings.

References

1. Ebrahim Bagheri, Faezeh Ensan, and Feras Al-Obeidat. Neural word and entity embeddings for ad hoc retrieval. *Information Processing & Management*,

54(4):657–673, 2018.
2. Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250, 2008.
3. Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 126–134, 2018.
4. Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.
5. David Guthrie, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. A closer look at skip-gram modelling. In *LREC*, volume 6, pages 1222–1225. Citeseer, 2006.
6. Siham Jabri, Azzeddine Dahbi, Taoufiq Gadi, and Abdelhak Bassir. Ranking of text documents using tf-idf weighting and association rules mining. In *2018 4th International Conference on Optimization and Applications (ICOA)*, pages 1–6. IEEE, 2018.
7. 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.
8. Zhenyang Li, Guangluan Xu, Xiao Liang, Feng Li, Lei Wang, and Daobing Zhang. Exploring the importance of entities in semantic ranking. *Information*, 10(2):39, 2019.
9. Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. *arXiv preprint arXiv:1805.07591*, 2018.
10. Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian. Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1101–1104, 2019.
11. Mandar Mitra and BB Chaudhuri. Information retrieval from documents: A survey. *Information retrieval*, 2(2-3):141–163, 2000.
12. Iqra Muhammad, Danushka Bollegala, Frans Coenen, Carol Gamble, Anna Kearney, and Paula Williamson. Maintaining curated document databases using a learning to rank model: The orrca experience. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 345–357. Springer, 2020.
13. Iqra Muhammad, Anna Kearney, Carrol Gamble, Frans Coenen, and Paula Williamson. Open information extraction for knowledge graph construction. In *International Conference on Database and Expert Systems Applications*, pages 103–113. Springer, 2020.
14. Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. Improving document ranking with dual word embeddings. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 83–84, 2016.
15. Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*, 2019.
16. 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.

17. Harshith Padigela, Hamed Zamani, and W Bruce Croft. Investigating the successes and failures of bert for passage re-ranking. *arXiv preprint arXiv:1905.01758*, 2019.
18. Jiaul H Paik. A novel tf-idf weighting scheme for effective ranking. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 343–352, 2013.
19. Karl Pearson. The problem of the random walk. *Nature*, 72(1867):342–342, 1905.
20. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
21. Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.
22. Thomas Rebele, Fabian Suchanek, Johannes Hoffart, Joanna Biega, Erdal Kuzey, and Gerhard Weikum. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In *International semantic web conference*, pages 177–185. Springer, 2016.
23. Xuan Shan, Chuanjie Liu, Yiqian Xia, Qi Chen, Yusi Zhang, Angen Luo, and Yuxiang Luo. Bison: Bm25-weighted self-attention framework for multi-fields document search. *arXiv preprint arXiv:2007.05186*, 2020.
24. Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, 2018.
25. Lindsay S Uman. Systematic reviews and meta-analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, 20(1):57, 2011.
26. Chenyan Xiong, Zhuyun Dai, Jamie Callan, Zhiyuan Liu, and Russell Power. End-to-end neural ad-hoc ranking with kernel pooling. In *Proceedings of the 40th International ACM SIGIR conference on research and development in information retrieval*, pages 55–64, 2017.
27. Chenyan Xiong, Russell Power, and Jamie Callan. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279, 2017.
28. Hamed Zamani and W Bruce Croft. Relevance-based word embedding. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 505–514, 2017.