IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 23, 2021, accepted October 6, 2021, date of publication October 15, 2021, date of current version November 8, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3120727

NeuSub: A Neural Submodular Approach for Citation Recommendation

BINH THANH KIEU[1,2], INIGO JAUREGI UNANUE[1,3], SON BAO PHAM[1,2], HIEU XUAN PHAN[2], AND MASSIMO PICCARDI[1], (Senior Member, IEEE)

[1]Faculty of Engineering and Information Technology, University of Technology Sydney, Broadway, NSW 2007, Australia
[2]Faculty of Information Technology, VNU University of Engineering and Technology, Hanoi 11309, Vietnam
[3]RoZetta Technology, Sydney, NSW 2000, Australia

Corresponding author: Binh Thanh Kieu (binhkt@vnu.edu.vn)

This work was supported in part by the Joint Technology and Innovation Research Centre—a partnership between the University of Technology Sydney and Vietnam National University.

ABSTRACT Citation recommendation is a task that aims to automatically select suitable references for a working manuscript. This task has become increasingly urgent as the typical pools of candidates continue to grow, in the order of tens or hundreds of thousands or more. While several approaches to citation recommendation have been proposed in the literature, they generally seem to lack principled mechanisms to ensure diversity and other global properties among the recommended citations. For this reason, in this paper we propose a novel citation recommendation approach that leverages a submodular scoring function and a deep document representation to achieve an effective trade-off between relevance to the query and diversity of the references. To optimally train the scoring function and the deep representation, we propose a novel training objective based on a structural/multiclass hinge loss and incremental recommendations. The experimental results over three popular citation datasets have showed that the proposed approach has led to remarkable accuracy improvements, with an increase of up to 1.91 pp of MRR and 3.29 pp of F1@100 score with respect to a state-of-the-art citation recommendation system.

INDEX TERMS Citation recommendation, deep neural networks, structural/multiclass hinge loss, submodular inference, transformer models, BERT, sentence-BERT.

I. INTRODUCTION

Citation recommendation is a popular task of natural language processing and information retrieval that aims to automate the selection of references for a working manuscript. Citation recommendation promises to be especially useful for investigators who are approaching a new topic or an unfamiliar field, as well as for researchers in-training, and it is becoming increasingly urgent as the size of the typical candidate pools continues to grow (in the order of tens or hundreds of thousands candidates or more). In a plausible scenario of use, the authors of a manuscript - for instance, a paper draft to be submitted to IEEE Access - may be already familiar with a few, key references, but would wish to turn to an automated tool to provide an exhaustive or supplementary list. The main goal of citation recommendation is to speed up and facilitate this selection.

A common approach to citation recommendation, called *local* or *context-aware* citation recommendation, is to select a short text of typically 1-3 sentences (approximately 50-100 words) as a query, and to identify the most suitable citation(s) for it. This approach has been followed by many works, including, among others, He *et al.* [1], Huang *et al.* [2], Jeong *et al.* [3], and, more recently, Färber and Sampath [4]. Another line of work, known as *global* citation recommendation, aims to instead identify all the most suitable references for a given manuscript as a single, overall recommendation. Much-cited approaches in this category include Ren *et al.* [5] and Dai *et al.* [6]. Since global citation recommendation does not preclude the possibility to exploit local searches, it can be regarded as a more general approach and we thus follow it in this work. In all cases, the recommendation approaches can take advantage of the *citation network* formed by the candidate documents (i.e., the graph of their mutual citations) and information such as their title, abstract and metadata, inclusive of authors, venues, publication dates and so forth.

The associate editor coordinating the review of this manuscript and approving it for publication was Imran Sarwar Bajwa.

In general, the key components of query-based citation recommendation systems are the representations chosen for the query and the documents, and the scoring functions used to assess the similarity between the query and the candidates, and between the candidates themselves.

Most existing citation recommendation systems rank the candidate documents based on their relevance to the given query, and recommend the top entries. While this approach is generally effective, it inherently lacks the ability to score inter-candidate properties such as, for instance, redundancy of content or diversity of authors. In alternative to simple relevance ranking, other approaches have therefore proposed using *submodular scoring functions* to select the best candidates based on trade-offs between relevance to the query and inter-candidate properties. For instance, Kieu *et al.* in [7] have proposed scoring the candidates based on a combination of relevance, coverage and diversity, while Yu *et al.* in [8] have proposed scoring the ''information flow'' of a candidate within the citation network. In addition to works that have addressed the limitations of the scoring function, other works have addressed the limitations of the document representation. For instance, Bhagavatula *et al.* in [9] have proposed representing the documents using compositional word embeddings, while Jeong *et al.* in [3] have proposed using contemporary transformer models. However, to the best of our knowledge no work has yet addressed the *integration* between submodular inference and trainable document representations. For this reason, in this paper we aim to fill this gap by presenting a novel training objective optimized for submodular selection. The main contributions of our work can be summarized as follows:

- a novel approach, nicknamed *NeuSub*, for scoring sets of candidate documents based on submodular inference and a deep document representation (Sections IV-B-C);
- an original training objective based on a structural/multiclass hinge loss for optimally training the deep document representation (Section IV-D);
- a comprehensive experimental comparison with competitive approaches, including strong baselines and a state-of-the-art model, on a number of probing citation datasets (the ACL Anthology Network, DBLP and PubMed). The experimental results give strong evidence to the superior accuracy of the proposed approach (Section V).

The rest of this paper is organized as follows: Section II discusses the main related work, while Section III introduces the problem formulation. Section IV presents the entire methodology, inclusive of the submodular scoring function, the document representation, and the submodularity-oriented training objective. Section V describes the experiments and the main results, and finally Section VI presents the conclusions.

II. RELATED WORK

From the perspective of the underlying technology, citation recommendation approaches can be divided into three main groups: 1) content-based approaches, 2) collaborative filtering approaches, 3) graph-based approaches, and 4) hybrid approaches. Content-based approaches first compute a similarity score between the user's query (for instance, a paper's draft, abstract, or topic) and the text from all the candidate documents, and then recommend the top K scoring documents as citations [9]. Collaborative filtering approaches address the problem of citation recommendation from a ''social'' perspective, by recommending similar citations to those of users with a similar profile. In a way, they shift the focus from document similarity to user similarity [10]. However, their main acknowledged limitation is that user similarity falls short of relevant recommendations for new authors and authors exploring new topics. Graph-based approaches leverage the structure of the citation network to derive embeddings for the nodes (the papers) and recommend citations as a task of edge, or link, prediction [11]–[18]. Due to their nature, only a subset of these approaches can provide predictions for new nodes. Finally, under hybrid approaches we categorize all the approaches that mix content-based and collaborative filtering techniques, exploiting the available citation network in various ways [4], [19], [20]. Another important distinction is between local approaches that only use a small, local window of the draft as the query (i.e., ''find the most suitable citation for this citation context'') and approaches with the more ambitious goal of recommending an overall citation list, globally optimal for the given draft [3], [4]. In our paper, we focus on the latter, and on content-based recommendation which seems to rely on the most general assumptions. In the following, we briefly review the existing approaches that are more immediately relevant to the proposed approach.

A. NEURAL NETWORK APPROACHES

Neural network approaches, and in particular deep learning-based, have established an impressive record of achievement over a variety of tasks and data, including images, text, and metadata. In more recent years, they have started to supplant more conventional methods also in citation recommendation [21]. In one of the earliest works, Huang *et al.* in [2] have proposed using a multilayer neural network to learn representations of words and documents, and compute the probability of recommending a document for a given citation context. Their results have showed an improvement of more than 9 percentage points of recall compared with a topic-based model. Gupta and Varma in [22] have addressed citation recommendation by introducing a novel approach that combines a document's neural embedding with a graph structure. Their results have showed that their approach has been able to outperform two state-of-the-art models (one based on TF-IDF representations and the other on latent Dirichlet allocation, or LDA) by an impressive 21 and 39 percentage points of mean average precision (MAP), respectively. Jeong *et al.* in [3] have used a combination of a BERT transformer model and a graph convolutional network to improve a local citation recommendation model. In this way, they have been able to leverage the representational power of pre-trained language models that have become a key component in a wide range of natural language processing tasks [23]. A recent deep learning-based approach from the Allen Institute for AI (AI2),[1] Citeomatic [9], has been able to establish a new state of the art for the field by neatly dividing the approach into two stages: in the first, the model selects a vastly abundant number of citations (e.g., $1,000$) for the given query document based on similarity of neural embeddings; in the second, it uses a trained neural reranker to rerank the selected documents and recommend the top K.

In general, it seems that approaches based on neural networks and deep learning have been able to outperform more traditional citation recommendation approaches, that heavily relied on fixed document representations (i.e., TF-IDF), topic descriptors (i.e., LDA, LSI etc), and clustering [5], [6], [24]. In addition, some of the existing approaches require access to the full text of the documents, which is often not available due to paywalled content, while others are able to operate solely on the abstracts and metadata, which are instead usually publicly available. Both our approach and the state-of-the-art Citeomatic [9] fall in this category, and we therefore use Citeomatic as our main term of reference in the experiments.

B. SUBMODULAR APPROACHES

Regardless of the nature of the approach, most previous works address citation recommendation as a ranking task based on relevance (or, even simply, similarity) to the query. This means that each document in a pool of candidates is individually scored based on the query, and the K top scoring documents are chosen as the citation recommendations. A more comprehensive line of attack makes use of submodular scoring functions for the selection: rather than simply selecting the top scoring documents, submodular approaches choose each recommendation incrementally, based not only on the candidates' relevance scores, but also on the set of citations already recommended. While more expensive computationally, submodular approaches can generate recommendation lists that are more jointly optimal; i.e., which consist of complementary, less-redundant citations that are selected in a manner more similar to that in which human experts carry out manual selections. Given the attractive properties of submodularity, a few submodular citation recommendation approaches have been proposed in the last few years [7], [8]. Yu *et al.* in [8] have used a submodular approach to optimize the "information flow" in a citation network. Kieu *et al.* in [7] have proposed a submodular approach based on combinations of relevance to the query, coverage of the corpus and diversity of the list. Their recommendation approach consists of two main components: 1) a document similarity scoring function, and 2) a submodular selection procedure. The document similarity function is used to compute a similarity score between any pair of documents (either the query and a candidate, or any two documents in the corpus). For submodular selection, they have explored a number of submodular functions, including monotone and non-monotone, and with and without meta-information. Differently from previous approaches, our approach leverages a deep textual representation of the documents, fine-tuned with a dedicated objective derived from the citation graph, and uses it for submodular selection of the recommended citations.

III. PROBLEM FORMULATION

A multi-label classification (MLC) task is a classification task where the object to be classified can belong to multiple, and possibly all, classes in a given class set. MLC tasks are commonplace in many domains and include applications as diverse as classification of proteins, categorization of music pieces, semantic classification of scenes, and so forth [25]. Global citation recommendation can be seen as an instance of multi-label classification, where the goal of the task is to assign a query document to a set of citations out of a typically large ($> 10,000$) set of candidates. The selection can leverage both the candidates' content (title, abstract, text) and metadata (authors, venue, publication year etc), where available.

For a formal description of a multi-label classification task, we denote the set of classes as $D = \{d_j : j = 1 \ldots N\}$ and a set of labelled instances as $\{(x_i, Y_i) : i = 1 \ldots M\}$, where x_i is the available vector of measurements for the i-th instance, and Y_i is its subset of labels. The label subset can be simply stored as a binary vector, $Y_i = \{y_1, \ldots y_N\}$, where every $y_j = 1$ indicates the presence of label d_j in the class set for x_i. Using this convention, the output space can be noted as $Y = \{0, 1\}^N$. The aim of an MLC task is to assign all and only the correct labels to any new measurement, x. In the case of citation recommendation, D is the set of candidate documents, x_i is the measurement vector for the i-th query document (typically designed as a vector of similarity measurements between the query, q, and the candidates in D), and Y_i is its ground-truth reference list. In turn, citation recommendation can be framed as the building of a scoring function that can score any subset of the candidates and allows choosing the best K as the recommended citations, where K is a "budget" typically chosen by the query's authors. For a given query, let us note a subset of the candidates as S, and the scoring function as $f(S)$. This function certainly depends on many other parameters, including: 1) the query document itself, q; 2) the pool of candidates, D; 3) the model used to represent the documents, with its own parameters; and 4) the function chosen to measure the pairwise similarity of any two given documents, with its own parameters. However, we leave all such dependencies implicit in the following to keep the notations concise. The problem of citation recommendation can thus be formally expressed as:

$$\begin{aligned} \bar{S} &= \text{argmax}_S f(S) \\ \text{s.t.} \quad & S \subset D, |S| = K \end{aligned} \tag{1}$$

[1] https://allenai.org/

Selecting the optimal subset of size K, $\bar{S}$, out of N candidate documents is an NP-hard problem with a prohibitive $O(\binom{N}{K})$ complexity. To understand the complexity more clearly, let us consider the case at hand, where the number of available of classes, N, is much larger than that of the allowable selections, K. In this case, $\binom{N}{K} \approx \frac{N^K}{K!}$ from Stirling's approximation.[2] In turn, $\frac{N^K}{K!} > \frac{N^K}{K^K} = (\frac{N}{K})^K$. For instance, for $N = 20,000$ and $K = 50$, this translates into a prohibitive $\approx 10^{130}$ evaluations to select the optimal subset. However, several approximate selection strategies are possible, starting from simply selecting the K documents that are the best individual singletons. In the following section, we discuss submodular approaches.

IV. METHODOLOGY

In our approach, we address the task of global citation recommendation by combining a deep neural representation of the query and candidate documents with a stage of submodular selection of the recommended citations. In this section, we describe the proposed approach, including the submodular inference (Section IV-A), the document representation and similarity measure (Section IV-B), and the submodularity-oriented training objective (Section IV-C). For convenience, in Table 1 we concisely present the main notations used in this paper.

TABLE 1. Main notations used in this paper (approximately in order of appearance).

Notation	Definition
D	The set of citable documents
N	The number of citable documents (i.e., the size of D)
Y	The ground-truth reference list of a query document (a subset of D)
x	The measurement vector of a query document
S	The predicted reference list for a query document (i.e., the citation recommendation, a subset of D)
K	The number of predicted references, or "budget" (i.e., the size of S)
$f()$	The scoring function of any subset of D (NB: higher scores are better)
A, B	Two auxiliary subsets of D
S_k	The set of the first k predicted references
d	A generic document $\in D$
$S_k \cup d$	The union set of S_k and d
q	The query document (wherever it needs to be explicitly mentioned)
$s(q, d)$	The pairwise similarity between q and d
$P_i, i = 1 \dots C$	The clusters of a partition of D
$\mathcal{L}$	The training loss function
Y_k	A subset of size k of the ground-truth reference list, Y
$d^g \in Y \setminus Y_k$	A document in the remaining ground-truth reference list, $Y \setminus Y_k$
$d \in D \setminus Y$	A document not in the ground-truth reference list
$d^* \in D \setminus Y$	The "most violating" document not in the ground-truth reference list
m	The margin parameter in the loss function
$\mathcal{L}^*$	The training loss function only accounting for d^*

A. SUBMODULAR INFERENCE

A scoring function such as $f(S)$ is said to be *submodular* if the following property holds:

$$[f(A \cup d) - f(A)] \geq [f(B \cup d) - f(B)] \quad \forall A \subseteq B \subseteq D \tag{2}$$

In Eq. 2, A and B are two subsets of documents, with A "smaller" than or equal to B, and $(A \cup d)$ and $(B \cup d)$ represent their respective union sets with an additional document, d. This property, known as the "law of diminishing returns", states that adding a new element, d, to a subset brings less benefit to the score the larger such subset is. In addition, if $A \subseteq B \rightarrow f(A) \leq f(B)$, the scoring function is said to be monotonic (i.e., adding a new element never decreases the score). An important result due to Nemhauser *et al.* [26] applies to monotonic submodular scoring functions: let us assume that the recommended set of K citations is selected with the following greedy algorithm: starting from an empty set, S_0, at iteration $k = 1 \dots K$, the algorithm adds document $d \in D \setminus S_{k-1}$ that maximizes $f(S_{k-1} \cup d)$:

$$S_k = S_{k-1} \cup \text{argmax}_{d \in D \setminus S_{k-1}} f(S_{k-1} \cup d), \quad k = 1 \dots K \tag{3}$$

The final inferred set, S_K, enjoys a lower bound on performance that ensures that $f(S_K)$ is at least $(e-1)/e$ of $f(S)$'s absolute, unknown maximum [26]. This result proves that the greedy inference algorithm is an efficient and effective approximation in the case of monotonic submodular functions. Note also that the K elements are chosen one at a time, but not independently. This affords the possibility to avoid redundancy, which would be impossible to enforce by choosing the K elements independently. For this reason, we have used submodular inference for both the selection of the citations and the training of our model. As scoring function, $f(S)$, we have adopted the function proposed in [7]:

$$f(S) = \sum_{i=1}^{C} \sqrt{\sum_{d \in (S \cap P_i)} s(q, d)} \tag{4}$$

where $P_i, i = 1 \dots C$, represent the clusters of a partition of the candidate documents, obtained by clustering the documents by either authors or venues, and $s(q, d) \geq 0$ is a pairwise similarity function between the query and a document. It is straightforward to prove that this function is monotonic (all terms are non-negative) and submodular (as a cluster grows larger, the square root reduces the benefit of adding a new document). In addition, the square root favors selecting citations from different clusters, increasing the diversity of the selection and mollifying the risk of redundancy. Differently from [7], instead of measuring the similarity with the fixed cosine similarity metric, we have used a trained probability of similarity provided by a deep learning module, described in the following subsection.

B. DOCUMENT REPRESENTATION AND SIMILARITY

An effective document representation is a key requirement for accurate citation recommendations. In addition, since citation recommendation inherently relies on comparisons between the query and the candidates, the representation has

[2] https://en.wikipedia.org/wiki/Stirling_approximation.

to also support efficient comparisons. Traditional document representations such as TF-IDF are efficient, but their general nature may limit their effectiveness compared to dedicated, learned representations. For this reason, in our approach we represent each document with a dense vector generated by a contemporary transformer model, BERT [23]. BERT is a powerful and flexible model that allows, among other, comparing any two text sequences provided in input. However, its comparison speed is rather limited, and for this reason Reimers and Gurevych in [27] have proposed Sentence-BERT, a BERT variant that makes text encoding and comparison faster by orders of magnitude. In our approach, we employ Sentence-BERT (or SBERT for short) to encode the query and the individual candidate documents. As underlying transformer, we have used DistilBERT/bert-base-uncased that has approximately 66M parameters. For a given document, we build the input by prepending a starting [*CLS*] token, followed by the document's title, a [*SEP*] (i.e., separator) token, and the document's abstract. The title and abstract are, in turn, encoded with WordPiece [28]. We then process the input with SBERT and retain the final layer's encoding of the [*CLS*] token as the overall representation of the document. In addition to the encoding, SBERT outputs the probability of similarity of the two input documents, and we use this probability as $s(q, d)$ in Eq. 4.

C. SUBMODULARITY-ORIENTED TRAINING OBJECTIVE

The goal of our training approach is to optimally train scoring function $f(S_{k-1} \cup d)$ in Eq. 3. This entails training f to make optimal, *incremental* decisions about the documents to add to the partial citation lists. To this aim, we first build a training set of (Y_k, d^g) pairs, where Y_k is a subset of size k of the query's ground-truth list, Y, and d^g is a document from the query's remaining ground-truth citations, $Y \setminus Y_k$. To limit the size of this training set, we bound the maximum number of elements in Y_k, but we generate all possible combinations of subsets and ground-truth elements within the maximum number.

However, training such an incremental function is very challenging because of the large space of combinations of possible subsets and additional ground-truth citations. For this reason, we resort to a two-step training strategy:

- in the first step, we pre-train the model using a standard Siamese network approach [29]. In this approach, only two classes are considered: documents from the query's ground truth, and other documents. The training loss is the conventional contrastive loss and the approach is useful to "warm-start" the model;
- as the second step, we propose an original training loss based on *structural/multiclass SVM* [30], using the purposely-created incremental subsets. This step fine-tunes the model to effectively support the incremental decisions.

The loss proposed for the second step is based on the structural/multiclass SVM framework [30] and can be expressed as:

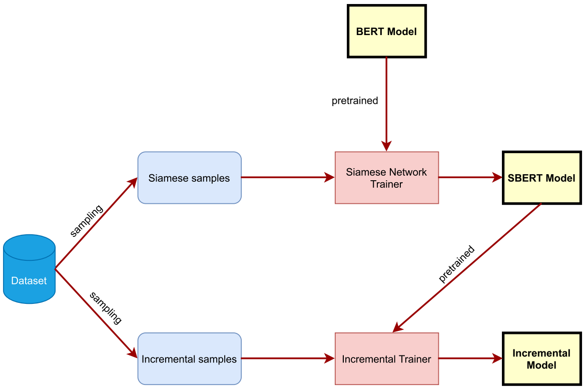

FIGURE 1. Overview of the proposed NeuSub training approach.

$$\mathcal{L} = \max[0, f(S_k \cup d) - f(S_k \cup d^g) + m], \quad d \in D \setminus Y \tag{5}$$

where d is a document that does not belong to the query's ground truth ($d \in D \setminus Y$), and m is a chosen constant. This loss function is equal to zero only if the score assigned to the ground-truth document, d^g, is larger than that assigned to the non-ground-truth document, d, by a margin of at least m. In any other case, the loss is > 0, and training will attempt to decrease it by the usual gradient descent and backpropagation through the model's parameters. However, the number of non-ground-truth documents is typically very large (i.e., tens or hundreds of thousands), and training may not afford to minimize a corresponding number of loss functions. Therefore, the approach only finds the largest of such losses by searching for the "most violating" non-ground-truth document:

$$d^* = \text{argmax}_{d \in D \setminus Y} \mathcal{L} \tag{6}$$

and including only its loss:

$$\mathcal{L}^* = \max[0, f(S_k \cup d^*) - f(S_k \cup d^g) + m] \tag{7}$$

in the minimization. The steps of the proposed training algorithm can thus be recapped as:

1) For the given query, compute the score of the current citation list plus the new ground-truth document, $f(S_k \cup d^g)$.
2) Find the most-violating non-ground-truth document, d^*, using Eq. 6. This entails evaluating $f(S_k \cup d)\ \forall d \in D \setminus Y$.
3) Form the loss function according to Eq. 7.
4) Using automatic differentiation, compute the gradient of the loss function and backpropagate it through the model.

For convenience, Figure 1 shows a high-level overview of the overall, proposed training approach, while Figure 2 shows an analogous overview of the inference stage.

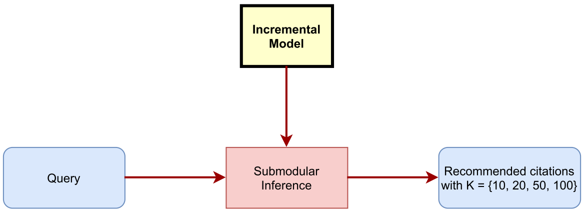

FIGURE 2. Overview of the proposed NeuSub inference.

TABLE 2. Main statistics of the citation datasets.

Dataset	AAN	DBLP	PubMed
# documents, training set	20,077	28,807	34,230
Year range, training set	1965-2013	1966-2007	1966-2008
# documents, validation set	663	773	2,820
Year range, validation set	2014-2014	2008-2008	2009-2009
# documents, test set	673	931	8,815
Year range, test set	2014-2014	2009-2010	2010-2013
total # samples	21,413	30,511	45,865

V. EXPERIMENTS AND RESULTS

A. DATASETS

For the experiments, we have used three probing citation datasets, namely the ACL Anthology Network corpus (AAN), DBLP and PubMed. Their main statistics are shown in Table 2.

The first dataset, **AAN**, is a dataset of papers published by the Association for Computational Linguistics between 1965 and 2014 and is a de-facto benchmark for the field. For the experiments, we have split it over the three sets: the papers from 1965 to 2013 as training set, approximately half randomly selected papers from 2014 as validation set, and the rest as test set. Choosing the validation and test sets from the same or similar years makes it more likely that they have comparable distributions. After filtering out the documents with fewer than 3 citations, the number of documents in the validation and test sets have become 663 and 673, respectively.

The other two datasets, **DBLP** and **PubMed**, are also popular datasets for citation recommendation evaluation. For these two datasets, we have retained the training, validation and test splits used by the state-of-the-art Citeomatic system [9] for a more direct comparison. For both these datasets, the documents with fewer than 10 citations have been filtered out from the validation and test sets.

It is important to note that for a fair evaluation, all the ground-truth references of each document need to be citable documents. For this reason, we have restricted the ground-truth reference lists of all the documents to only the documents within the relevant training set. For some documents, this restriction has led to empty ground-truth reference lists; we have retained these documents in the training sets, yet excluded them from the validation and test sets. Table 3 shows the ranges and average numbers of ground-truth references per document in the datasets and the splits used for the experiments. The average values for AAN and DBLP are roughly comparable, while those of PubMed are much higher.

TABLE 3. Minimum, maximum and average number of citations per document in the citation datasets.

Dataset	AAN	DBLP	PubMed
Min-max # references, training set	0-88	0-78	0-275
Average # references, training set	5.46	3.68	17.59
Min-max # references, validation set	3-40	10-56	10-157
Average # references, validation set	9.98	13.21	24.90
Min-max # references, test set	3-46	10-83	10-286
Average # references, test set	9.92	15.17	31.60

B. EVALUATION METRICS

Several metrics have been in common use for evaluating the performance of citation recommendation systems, including the mean reciprocal rank (MRR), the mean average precision (MAP), and the normalized discounted cumulative gain (NDCG) [19]. To provide a detailed view of the performance, a very informative and comprehensive metric is the F1 score at different levels of recommendation (the number of recommended citations, or "budget", K). In addition, the F1 score can be computed at either "micro" or "macro" level. In the micro F1 score, the statistics of correct and incorrect predictions are accumulated over the entire corpus first, and then a single, final F1 score is computed. In the macro-F1 score, instead, the statistics are accumulated over the individual documents, and their F1 scores are then computed and averaged. Since the ground-truth reference lists of the individual documents are very different in length, weighing all the documents equally with the macro averaging does not seem appropriate for assessing this task. Therefore, we have used the micro F1 score "at K" to evaluate the effectiveness of the compared models, with budgets of $K = 10$, 20, 50 and 100 elements. The micro F1 score with budget $K = 20$ was also the main measure used for the evaluation of Citeomatic in [9]. The micro F1 score for a K budget, F1@K, can be simply expressed as:

$$\begin{aligned} \text{F1@}K &= 2 * \frac{\text{prec@}K * \text{rec@K}}{\text{prec@}K + \text{rec@K}} \\ \text{prec@}K &= \frac{TP}{TP + FP} \\ \text{rec@}K &= \frac{TP}{TP + FN} \end{aligned} \tag{8}$$

where TP is the total number of correctly recommended citations over the entire set of documents, FP is the total number of incorrectly recommended citations, and FN is the total number of missed ground-truth citations, all from K predicted citations per document; prec@K and rec@K are the precision and recall at K.

Following [9], in addition to the F1@K scores we also report the mean reciprocal rank (MRR) of the best correct prediction of the various approaches. As applied by [9], the MRR measures the average of the reciprocal of the ranking position of the best correct prediction. By noting the ranking position as n, its reciprocal can be noted as $1/n$. For accurate recommendations, it is desirable that the best correct prediction be ranked high in prediction order (i.e., a small n) and, as a consequence, its reciprocal will be a relatively large value. The MRR reports the average of the reciprocal rank of the best correct predictions of all the test queries (range: 0 - worst, 1 - best).

C. COMPARED APPROACHES

We have compared the proposed approach with a strong baseline, a state-of-the-art model, and two existing algorithms that can be regarded as ablated versions of the proposed approach. The compared algorithms are:

- **Elasticsearch** with the Okapi BM25 similarity score function: Elasticsearch is a highly popular document retrieval algorithm, with a market share of 82.75% of the hosted-search market.[3] Elasticsearch uses the Okapi BM25 similarity score function which has proven superior to TF-IDF-based similarity in a number of benchmarks [31]. BM25 has two hyperparameters, a bias, b, and a scale, k_1 which have both been kept to their default values (0.75 and 1.2, respectively) in the experiments.
- **Citeomatic**: Citeomatic is a citation recommendation system developed by the Allen Institute for AI (AI2) that has established state-of-the-art results on all tested datasets. Citeomatic is a two-step approach that uses a neural embedding of the documents to provide an initial selection of the most similar documents to the query (e.g., $1,000$), followed by a reranking step that delivers the final K recommendations. For the experiments, we have used the code publicly released by the authors.[4]
- **SubRef**: SubRef is a citation recommendation algorithm that uses a submodular selection function comparable to that used in the proposed approach, and the BM25 similarity score to measure the pairwise document similarity.
- **SBERT**: SBERT is a citation recommendation algorithm that uses a submodular selection function comparable to that used in the proposed approach, and a Sentence-BERT module to learn a deep representation of the documents. However, the deep representation is learned using only the Siamese configuration with a heuristic selection of the negative ground-truth documents.

In addition to the above algorithms, in Section V-E we also provide a separate comparison with a state-of-the-art graph-based citation recommendation approach, **attri2vec** [13], in order to analyze and contrast the benefits and limitations of these two styles of approaches.

For training the proposed approach, NeuSub, we have first applied Siamese pre-training by following precisely the procedure presented in [32], with ground-truth negative samples chosen with the *farthest* heuristic and a citation distance $d = 2$. After completing the pre-training, we have trained the pre-trained model with our submodularity-oriented training objective, setting the maximum number of elements in the Y_k partial reference lists to 5. The model has been trained for 10 epochs in the case of the AAN dataset, while for the larger DBLP and PubMed it has been trained for 5 and 3 epochs, respectively, for an approximate parity of total training time. Table 4 reports all other main hyperparameters used for training the proposed approach. For validation, we have run an evaluation over the validation set every $50,000$ batches. Out of all the evaluated models, we have retained that with the highest F1@20 score, and used it blindly on the test set.

TABLE 4. Main hyperparameters used for training the proposed approach (NeuSub).

Hyperparameter	Value
Batch size	16
Learning rate	0.001
L2 regularization	1e-5
L1 regularization	1e-7
Dropout rate	0.1
Word embedding dimension	768
Margin parameter	1.0
BERT model	DistilBERT
Vocabulary size	30,522
Optimizer	LazyAdamOptimizer

TABLE 5. Results on the AAN test set.

	MRR	F1@10	F1@20	F1@50	F1@100
Elasticsearch	0.2154	0.1096	0.1182	0.0784	0.0421
Citeomatic					
select	0.2954	0.1281	0.1340	0.0940	0.0548
select + rerank	0.3214	0.1395	0.1462	0.1042	0.0612
SubRef (best)	0.2786	0.1204	0.1287	0.0898	0.0501
SBERT (best)					
base	0.2843	0.1276	0.1347	0.0934	0.0549
base + submod	0.3067	0.1349	0.1461	0.1074	0.0637
NeuSub					
base	0.3095	0.1330	0.1429	0.1017	0.0602
base + submod	**0.3298**	**0.1516**	**0.1601**	**0.1199**	**0.0696**

D. RESULTS

Table 5 shows the results for all the compared models over the AAN test set. For SBERT and NeuSub, we report both the results without the submodular selection (*base*, i.e. only the top-K most similar documents to the query) and with it (*base* + *submod*). For Citeomatic, we report both the results for the K most similar documents from the initial selection (*select*), and after reranking (*select* + *rerank*). Table 5 shows that Elasticsearch has achieved the least accurate results of all the compared models. Conversely, the second-best results have been obtained by Citeomatic (select + rerank) for MRR, F1@10 and F1@20, and by SBERT (base + submod) for F1@50 and F1@100. The proposed approach, NeuSub (base + submod), has obtained the best MRR and the best F1 scores at all levels of budget, with improvements of 1.21 pp (percentage points) over Citeomatic in F1@10, and of 1.39 pp in F1@20.

Table 6 shows the main results over the DBLP test set. In this case, the MRR and the F1 scores are much higher in absolute value, but the relative rankings across the compared models remain similar. For this dataset, the second-best results for all budgets have been achieved by

[3] https://www.slintel.com/tech/hosted-search/elasticsearch-market-share
[4] https://github.com/allenai/citeomatic

TABLE 6. Results on the DBLP test set.

	MRR	F1@10	F1@20	F1@50	F1@100
Elasticsearch	0.3718	0.1696	0.1686	0.1247	0.1028
Citeomatic					
select	0.5790	0.3028	0.2820	0.1569	0.1478
select + rerank	0.6720	0.3128	0.3030	0.1656	0.1545
SubRef (best)	0.4518	0.2402	0.2296	0.1434	0.1301
SBERT (best)					
base	0.5401	0.2810	0.2743	0.1532	0.1460
base + submod	0.6412	0.3110	0.2943	0.1632	0.1510
NeuSub					
base	0.6027	0.2951	0.2901	0.1634	0.1510
base + submod	**0.6911**	**0.3213**	**0.3054**	**0.1687**	**0.1568**

TABLE 7. Results on the PubMed test set.

	MRR	F1@10	F1@20	F1@50	F1@100
Elasticsearch	0.2126	0.0748	0.1493	0.1385	0.1251
Citeomatic					
select	0.6790	0.2317	0.2842	0.2667	0.1946
select + rerank	**0.7054**	**0.2670**	**0.3092**	0.2768	0.1970
SubRef (best)	0.4682	0.1263	0.1971	0.1950	0.1724
SBERT (best)					
base	0.5579	0.1767	0.2686	0.2512	0.1878
base + submod	0.6217	0.1905	0.2825	0.2878	0.2092
NeuSub					
base	0.5931	0.1850	0.2781	0.2678	0.1978
base + submod	0.6501	0.2075	0.3028	**0.2985**	**0.2299**

Citeomatic, while the best results have all been achieved by the proposed approach. The improvements of NeuSub vs Citeomatic have ranged from 0.24 pp in F1@20 to 1.91 pp in MRR. It is interesting to note that the submodular inference of NeuSub has played a major role in its performance, with an improvement of 2.62 F1@10 pp compared to the same model without submodular selection.

Finally, Table 7 shows the main results on the PubMed test set. In this case, NeuSub has been outperformed by Citeomatic in MRR, F1@10 and F1@20, but has achieved the best scores in F1@50 and F1@100. The improvement in F1@100 has been particularly impressive, with an increase of 3.29 pp with respect to Citeomatic. Once again, NeuSub's submodular inference has played a key role in its performance, with improvements of up to 3.21 pp compared to the same models without submodular selection.

In addition to the quantitative results, in Table 8 we show a qualitative example from the AAN dataset. The query (paper of ID P14-1074 in the dataset) is titled: "Linguistic Structured Sparsity in Text Categorization" and has 13 ground-truth references, whose titles and IDs are also listed in the table. For conciseness, the table only reports the true positive predictions at $K = 100$ from Citeomatic and from the proposed approach, NeuSub. For this example, NeuSub has been able to retrieve 5 correct citations, while Citeomatic has retrieved only 3. In addition, NeuSub seems to have been able to retrieve some "hard-to-find" citations such as: "Predicting a Scientific Community's Response to an Article" (ID D11-1055) and: "Predicting Risk from Financial Reports with Regression" (ID N09-1031), that do not have any obvious similarity to the query's title.

Overall, submodularity has given evidence to be a strong property to leverage in citation selection. Its main advantage over a generic scoring function is in its intrinsic ability to foster diversity among the recommended citations, rather than only relevance to the query. We speculate that this somehow reflects how authors select citations for their own documents, and for this reason submodular selection has been able to better match human-generated citation lists. Another notable advantage of submodular selection is that it can work in tandem with human recommendations: a human expert can first manually choose a few citations that they are confident of, and submodular selection can then choose the remaining citations automatically, taking into accounts the citations already included by the expert. This significantly expands the possible scenarios of utilization.

E. COMPARISON WITH GRAPH EMBEDDING APPROACHES

Another vein of approaches for data that form graphs, such as citation networks, are those based on *graph embeddings* [11]–[16]. Graph embeddings simultaneously analyze the topology of the graph (i.e., the edges between nodes) and the attributes of each node and edge to infer embeddings that can be used for tasks such as node classification and edge prediction. In our case, we can leverage graph embeddings and edge prediction to predict the references outgoing from the given query nodes. Graph embeddings come in two fundamentally different styles: *transductive* and *inductive*. In the transductive style, all the nodes of the graph are assumed to be provided at once, and they are used during both the training and test stages of the predictive tasks. Well-known examples of transductive graph embeddings include node2vec, TransE and GCN [11], [12], [15]. The inductive style is instead much more flexible as it allows making predictions also for new nodes that were not part of the initial graph (out-of-sample predictions). A task of citation recommendation requires the inductive settings, since its main stated goal is to recommend citations for newly-written documents. For this reason, we have conducted an experiment with a state-of-the-art inductive graph embedding approach, attri2vec [13]. Attri2vec learns a mapping of node attributes that is simultaneously informed by the graph structure and able to embed new nodes. It has been reported as outperforming graph embedding approaches such as node2vec, GraphSAGE and GCN in a citation recommendation task [33]. For the experiment, we have used its StellarGraph implementation [33], setting the various hyperparameters (number of walks, length of walks etc) so as to limit the total training time to approximately one day per dataset. For a fair comparison, we have used the same node attributes used by the proposed approach, NeuSub (i.e., the Sentence-BERT deep representations). We have also attempted to use attri2vec's default node attributes (a simplified BoW representation), but results have been generally worse. Table 9 shows the results for attri2vec vis-à-vis those of the best NeuSub configuration over the AAN, DBLP and PubMed test sets. The differences are remarkable and require contextualization: graph embedding

TABLE 8. An example from the AAN dataset: top: query (paper ID and title); two leftmost columns: ground-truth reference list of the query (paper ID and title); two rightmost columns: Citeomatic's and NeuSub's true predictions.

Paper ID	Paper Title	Citeomatic	NeuSub
Query (P14-1074)	**Linguistic Structured Sparsity in Text Categorization**		
Citations			
D10-1102	Multi-Level Structured Models for Document-Level Sentiment Classification	✓	✗
D11-1055	Predicting a Scientific Community's Response to an Article	✗	✓
D11-1139	Structured Sparsity in Structured Prediction	✗	✗
D13-1024	Structured Penalties for Log-Linear Language Models	✗	✗
D13-1170	Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank	✗	✗
J92-4003	Class-Based n-gram Models of Natural Language	✓	✓
N07-1051	Improved Inference for Unlexicalized Parsing	✗	✗
N09-1031	Predicting Risk from Financial Reports with Regression	✗	✓
N10-1038	Movie Reviews and Revenues: An Experiment in Text Regression	✓	✓
N12-1097	Textual Predictors of Bill Survival in Congressional Committees	✗	✗
P11-1137	Discovering Sociolinguistic Associations with Structured Sparsity	✗	✓
W06-1639	Get out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts	✗	✗
W09-3607	The ACL Anthology Network	✗	✗

TABLE 9. Comparison of attri2vec and NeuSub on the AAN, DBLP and PubMed test sets.

	MRR	F1@10	F1@20	F1@50	F1@100
AAN					
Attri2vec	0.1813	0.1038	0.1092	0.0977	0.0527
NeuSub	0.3298	0.1516	0.1601	0.1199	0.0696
DBLP					
Attri2vec	0.4045	0.1841	0.1729	0.1377	0.1156
NeuSub	0.6911	0.3213	0.3054	0.1687	0.1568
PubMed					
Attri2vec	0.3602	0.1233	0.1427	0.1214	0.1309
NeuSub	0.6501	0.2075	0.3028	0.2985	0.2299

approaches are typically evaluated in a transductive scenario, where the negative edges used for testing are simply a sample of the negative edges that have been excluded from the training stage. Instead, in the inductive settings, the query is an altogether new node, none of its negative edges have been seen during training, and it is tested against all of them. For this reason, the inductive evaluation is intrinsically much more probing. On the other hand, graph embedding approaches have complementary advantages over pure citation recommendation approaches, in that they can embed and predict multiple, heterogeneous types of edges (not only "x cites y"). It could be argued that dedicated citation recommendation approaches such as the proposed approach or Citeomatic [9] are more accurate on the specific task, while graph embedding approaches are more flexible and versatile by design.

VI. CONCLUSION

In this work, we have proposed NeuSub, a novel approach for citation recommendation based on a deep representation of the documents and a submodular inference function. The main novelty of the proposed approach is the training procedure of the deep representation that is based on an incremental scoring function aimed to mirror the submodular inference. The experimental results over three probing citation datasets (AAN, DBLP and PubMed) have showed that the proposed approach has been able to outperform all compared approaches, including a state-of-the-art, content-based model, Citeomatic, by a significant margin of F1 score for all budgets for the first two datasets, and in F1@50 and F1@100 for the third. In addition, submodularity has proved to be a key component of the approach, with improvements in F1@100 score of up to 3.21 percentage points. In the near future, we aim to expand the training of the incremental scoring function to larger partial lists. Yet, we will have to explore ways to mitigate the corresponding computational complexity. Another possible extension is the integration of the proposed global citation recommendation approach with localized citations in the text, to fully automate both reference selection and citation insertion.

REFERENCES

[1] Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in *Proc. 19th Int. Conf. World Wide Web*, Apr. 2010, pp. 421–430.

[2] W. Huang, Z. Wu, C. Liang, P. Mitra, and C. L. Giles, "A neural probabilistic model for context based citation recommendation," in *Proc. 29th AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2015, pp. 2404–2410.

[3] C. Jeong, S. Jang, E. Park, and S. Choi, "A context-aware citation recommendation model with BERT and graph convolutional networks," *Scientometrics*, vol. 124, no. 3, pp. 1907–1922, Sep. 2020.

[4] M. Färber and A. Sampath, "HybridCite: A hybrid model for context-aware citation recommendation," *CoRR*, vol. abs/2002.06406, pp. 1–10, Feb. 2020.

[5] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han, "ClusCite: Effective citation recommendation by information network-based clustering," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Aug. 2014, pp. 821–830.

[6] T. Dai, L. Zhu, Y. Wang, H. Zhang, X. Cai, and Y. Zheng, "Joint model feature regression and topic learning for global citation recommendation," *IEEE Access*, vol. 7, pp. 1706–1720, 2019.

[7] T. B. Kieu, B. S. Pham, X. H. Phan, and M. Piccardi, "A submodular approach for reference recommendation," in *Proc. 16th Int. Conf. Pacific Assoc. Comput. Linguistics (PACLING)*, 2019, pp. 3–14.

[8] Q. Yu, E. L. Xu, and S. Cui, "Submodular maximization with multi-knapsack constraints and its applications in scientific literature recommendations," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Dec. 2016, pp. 1295–1299.

[9] C. Bhagavatula, S. Feldman, R. Power, and W. Ammar, "Content-based citation recommendation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2018, pp. 238–251.

[10] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Commun. ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[11] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2013, pp. 2787–2795.

[12] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, and R. Rastogi, Eds., 2016, pp. 855–864.

[13] D. Zhang, J. Yin, X. Zhu, and C. Zhang, "Attributed network embedding via subspace discovery," *Data Mining Knowl. Discovery*, vol. 33, no. 6, pp. 1953–1980, Nov. 2019.

[14] W. L. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. 30th Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 1024–1034.

[15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

[16] C. Pornprasit, X. Liu, N. Kertkeidkachorn, K.-S. Kim, T. Noraset, and S. Tuarob, "ConvCN: A CNN-based citation network embedding algorithm towards citation recommendation," in *Proc. ACM/IEEE Joint Conf. Digit. Libraries*, Aug. 2020, pp. 433–436.

[17] F. Liu, Z. Cheng, L. Zhu, C. Liu, and L. Nie, "A2-GCN: An attribute-aware attentive GCN model for recommendation," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 26, 2020, 10.1109/TKDE.2020.3040772.

[18] J. Yu, H. Yin, J. Li, M. Gao, Z. Huang, and L. Cui, "Enhance social recommendation with adversarial graph convolutional networks," *IEEE Trans. Knowl. Data Eng.*, early access, Oct. 26, 2020, 10.1109/TKDE.2020.3033673.

[19] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia, "Scientific paper recommendation: A survey," *IEEE Access*, vol. 7, pp. 9324–9339, 2019.

[20] J. Chen, Y. Liu, S. Zhao, and Y. Zhang, "Citation recommendation based on weighted heterogeneous information network containing semantic linking," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 31–36.

[21] M. Färber and A. Jatowt, "Citation recommendation: Approaches and datasets," *Int. J. Digit. Libraries*, vol. 21, no. 4, pp. 375–405, Dec. 2020.

[22] S. Gupta and V. Varma, "Scientific article recommendation by using distributed representations of text and graph," in *Proc. 26th Int. Conf. World Wide Web Companion*, 2017, pp. 1267–1268.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol. (NAACL-HLT)*, 2019, pp. 4171–4186.

[24] N. Sakib, R. B. Ahmad, and K. Haruna, "A collaborative approach toward scientific paper recommendation using citation context," *IEEE Access*, vol. 8, pp. 51246–51255, 2020.

[25] D. Carrillo, V. F. López, and M. N. Moreno, "Multi-label classification for recommender systems," in *Trends in Practical Applications of Agents and Multiagent Systems*. Cham, Switzerland: Springer, 2013, pp. 181–188.

[26] G. Nemhauser, L. Wolsey, and M. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Programm.*, vol. 14, no. 1, pp. 265–294, 1978.

[27] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992.

[28] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, pp. 1–23, Sep. 2016. [Online]. Available: http://arxiv.org/abs/1609.08144

[29] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 1994, pp. 737–744.

[30] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, Sep. 2005.

[31] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, 2009.

[32] T. B. Kieu, J. U. Inigo, B. S. Pham, X. H. Phan, and M. Piccardi, "Learning neural textual representations for citation recommendation," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, 2021, pp. 4145–4152.

[33] StellarGraph. *Comparison of Link Prediction With Random Walks Based Node Embedding—StellarGraph 1.2.1 Documentation*. GitHub Repository. Accessed: 2018. [Online]. Available: https://stellargraph.readthedocs.io/en/stable/demos/linkprediction/homogeneous-comparison-link-prediction.html

BINH THANH KIEU received the B.Sc. degree in information technology and the M.Sc. degree in computer science from the University of Engineering and Technology (UET), Vietnam National University (VNU), Hanoi, in 2010 and 2015, respectively. He is currently pursuing the Ph.D. degree (dual doctoral program) with the University of Technology Sydney (UTS) and VNU-UET, Hanoi. He had been working as a Lecturer and a Researcher at VNU-UET, from 2014 to 2019. His main interests include information retrieval, recommender systems, natural language processing, and machine learning.

INIGO JAUREGI UNANUE received the B.Eng. degree in telecommunication systems from the University of Navarra, Donostia-San Sebastian, Spain, in 2016, and the Ph.D. degree in natural language processing from the University of Technology Sydney, in 2020. From 2014 to 2016, he was a Research Assistant at the Centro de Estudio e Investigaciones Tecnicas (CEIT). He is currently a Natural Language Processing and Machine Learning Researcher at RoZetta Technology, Sydney, NSW, Australia. His research interests include machine translation and low-resource natural language processing.

SON BAO PHAM received the Bachelor of Computer Science degree (Hons.) and the Ph.D. degree in computer science and engineering from UNSW. He is currently an Associate Professor with the Faculty of Information Technology, VNU University of Engineering and Technology. His research interests include natural language processing, knowledge acquisition, and artificial intelligence. He received the University Medal from UNSW.

HIEU XUAN PHAN received the B.S. and M.S. degrees in information technology and computer science from the College of Technology (Coltech, now UET), Vietnam National University in Hanoi (VNUH), in 2001 and 2003, respectively, and the Ph.D. degree in computer and information science from the Graduate School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), in 2006. From 2006 to 2008, he was a JSPS Postdoctoral Fellow at the Graduate School of Information Science (GSIS), Tohoku University. From 2008 to 2010, he was also a Research Fellow at the Centre for Health Informatics (CHI), University of New South Wales (UNSW), Sydney, NSW, Australia. He is currently an Associate Professor at VNUH-UET. His research interests include natural language processing, machine learning, information retrieval, web and text mining, and business intelligence.

MASSIMO PICCARDI (Senior Member, IEEE) received the M.Eng. and Ph.D. degrees from the University of Bologna, Bologna, Italy, in 1991 and 1995, respectively. He is currently a Full Professor of computer systems with the University of Technology Sydney, Australia. His research interests include natural language processing, computer vision, and pattern recognition. He has coauthored over 150 articles in these areas. He is a member of the IEEE Computer and Systems, Man, and Cybernetics Societies, and a member of the International Association for Pattern Recognition, and serves as an Associate Editor for the IEEE TRANSACTIONS ON BIG DATA.