Importance Assessment in Scholarly Networks

Saurav Manchanda and George Karypis
University of Minnesota, Twin Cities, USA
{manch043, karypis}@umn.edu

Abstract

We present approaches to estimate content-aware bibliometrics to quantitatively measure the scholarly impact of a publication. Traditional measures to assess quality-related aspects such as citation counts and h-index, do not take into account the content of the publications, which limits their ability to provide rigorous quality-related metrics and can significantly skew the results. Our proposed metric, denoted by Content Informed Index (CII), uses the content of the paper as a source of distant-supervision, to weight the edges of a citation network. These content-aware weights quantify the information in the citation i.e., these weights quantify the extent to which the cited-node informs the citing-node. The weights convert the original unweighted citation network to a weighted one. Consequently, this weighted network can be used to derive impact metrics for the various entities involved, like the publications, authors etc. We evaluate the weights estimated by our approach on three manually annotated datasets, where the annotations quantify the extent of information in the citation. Particularly, we evaluate how well the ranking imposed by our approach associates with the ranking imposed by the manual annotations. The proposed approach achieves up to 103% improvement in performance as compared to second best performing approach.

1 Introduction

Scientific, engineering, and technological (SET) innovations have been the drivers behind many of the significant positive advances in our modern economy, society, and life. To measure various impact-related aspects of these innovations various quantitative metrics have been developed and deployed. These metrics play an important role as they are used to influence how resources are allocated, assess the performance of personnel, identify intellectual property (IP)-related takeover targets, value a company's intangible assets (IP is such an asset), and identify strategic and/or emerging competitors.

Citation networks of peered-reviewed scholarly publications (e.g., journal/conference articles and patents) have widely been used and studied in order to derive such metrics for the various entities involved (e.g., articles, researchers, institutions, companies, journals, conferences, countries, etc. (Aguinis et al. 2012)). However, most of these traditional metrics, such as citation counts and h-index treat all citations and publications equally, and do not take into account the content of the publications and the context in which a prior scholarly work was cited. Another related line of work, such as PageRank (Page et al. 1999) and HITS (Kleinberg 1999) takes the node centrality into consideration (as a proxy for publication influence), but still operate in an content-agnostic manner. These content-agnostic metrics fail to reliably measure the scholarly impact of an article as they do not differentiate between the possible reasons a scholarly work is being cited. Being content-agnostic, these metrics can be easily manipulated by the presence of malicious entities, such as publication venues indulging in self-citations, which leads to high impact factor, or a group of scholars citing each others' work. For example, Journal Citation Reports (JCR)[1] routinely suppresses many journals that indulge in *citation stacking*, a practice where the reviewers and journal editors pressure authors to cite papers that either they wrote or that are published in "their" journal. Thus, there is a need to establish content-aware metrics to accurately and quantitatively measure various innovation-related aspects such as their significance, novelty, impact, and market value. Such metrics are essential for ensuring that SET-driven innovations will play an ever more significant role in the future.

In this paper, we propose machine-learning-driven approaches, that automatically estimate the weights of the edges in a citation network, such that edges with higher weights correspond to higher-impact citations. There has been considerable effort in the past to identify important citations (Valenzuela, Ha, and Etzioni 2015; Jurgens et al. 2018; Cohan et al. 2019). These approaches treat this task as a supervised text-classification problem, and thus, require the availability of training data with ground truth annotations. However, generating such labeled data is difficult and time consuming, especially when the meaning of the labels is user-defined. In contrast, our approaches are distant supervised, that require no manual annotation. The proposed approaches leverage the readily available content of the papers as a source of distant-supervision. Specifically, we for-

[1] http://help.incites.clarivate.com/incitesLiveJCR/JCRGroup/titleSuppressions.html

mulate the problem as how well the linear combination of the representations of the cited publication explains the representation of the citing publication. The weights in this linear-combination quantify the extent to which the cited-publication informs the citing-publication. We evaluate the weights estimated by our approach on three manually annotated datasets, where the annotations quantify the extent of information in the citation. Particularly, we evaluate how well the ranking imposed by our approach associates with the ranking imposed by the manual annotations. The proposed approach achieves up to 103% improvement in performance as compared to second best performing approach.

While our discussion and evaluation focuses on identifying informing citations, our approach is not restricted to this domain, and can be used to derive impact metrics for the various involved entities. For example, the content-aware weights estimated by the proposed approach convert the original unweighted citation network to a weighted one. Consequently, this weighted network can be used to derive impact metrics for the various involved entities, like the publications, authors etc. For example, to find the impact of a publication, the sum of weights outgoing from its corresponding node can be used to quantify the impact of the publication, instead of using vanilla citation count.

The reminder of the paper is organized as follows. Section 2 presents the related literature review. The paper discusses the proposed method in Section 3 followed by the experiments in Section 4. Section 5 discusses the results. Finally, Section 6 corresponds to the conclusions.

2 Related Work

The research areas relevant to the work present in this paper belong to *citation indexing*, *citation recommendation*, *link prediction* approaches, *distant-supervised credit attribution* approaches and *citation-intent classification* approaches. We briefly discuss these areas below:

Citation Indexing

A citation index indexes the links between publications that authors make when they cite other publications. Citation indexes aim to improve the dissemination and retrieval of scientific literature. CiteSeer (Giles, Bollacker, and Lawrence 1998; Li et al. 2006) is a first automated citation indexing system that works by downloading publications from the Web and converting them to text. It then parses the papers to extract the citations and the context in which the citations are made in the body of the paper, storing this information in a database. Other examples of popular citation indices include Google Scholar[2], Web of Science[3] by Clarivate Analytics, Scopus[4] by Elsevier and Semantic Scholar[5]. Some examples of subject-specific citation indices include INSPIRE-HEP[6] which covers high energy physics, PubMed[7], which covers life sciences and biomedical topics, and Astrophysics Data System[8] which covers astronomy and physics.

Citation recommendation

Citation recommendation describes the task of recommending citations for a given text. It is an essential task, as all claims written by the authors need to be backed up in order to ensure reliability and truthfulness. The approaches developed for citation recommendation can be grouped into 4 groups as follows(Färber and Jatowt 2020): hand-crafted feature based approaches, topic-modelling based approaches, machine-translation based approaches, and neural-network based approaches. Hand-crafted feature based approaches are based on features are are manually engineered by the developers. For example, text similarity between the citation context and the candidate papers can be used as one of the text-based features. Examples of papers that propose hand-crafted feature based approaches include (Färber and Jatowt 2020; He et al. 2011; LIU, YAN, and YAN 2016; Livne et al. 2014; Rokach et al. 1978). Topic modeling based approaches represent the candidate papers' text and the citation contexts by means of abstract topics, and thereby exploiting the latent semantic structure of texts. Examples of topic modeling based approaches include (He et al. 2010; Kataria, Mitra, and Bhatia 2010). The machine-translation based approaches apply the idea of translating the citation context into the cited document to find the candidate-papers worth citing. Examples in this category include (He et al. 2012; Huang et al. 2012). Finally, the popular examples of neural-network based models include (Ebesu and Fang 2017; Han et al. 2018; Huang et al. 2015; Kobayashi, Shimbo, and Matsumoto 2018; Tang, Wan, and Zhang 2014; Yin and Li 2017).

Link-prediction

A link is a connection between two nodes in a network. As such, link-prediction is the problem of predicting the existence of a link between two nodes in a network. A good link-prediction model predicts the likelihood of a link between two nodes, so it can not only be used to predict new links, but to also curate the graph by filtering less-likely links that are already present. Thus, the link-prediction can be a useful tool to find likely citations in a citation network. The citation recommendation task described previously can be thought of as a special case of link-prediction. Following the taxonomy described in (Martínez, Berzal, and Cubero 2016), link-prediction approaches can be broadly categorized into three categories: similarity-based approaches, probabilistic and statistical approaches and algorithmic approaches. The similarity based approaches assume that nodes tend to form links with other similar nodes, and that two nodes are similar if they are connected to similar nodes or are near in the network according to a given similarity function. Examples of popular similarity functions include number of common neighbors (Liben-Nowell and Kleinberg 2007), Adamic-Adar index (Adamic and Adar

[2] https://scholar.google.com/

[3] http://www.webofknowledge.com/

[4] https://www.scopus.com/

[5] https://www.semanticscholar.org/

[6] https://inspirehep.net/

[7] https://pubmed.ncbi.nlm.nih.gov/

[8] http://ads.harvard.edu/

2003), etc. The probabilistic and statistical approaches assume that the network has a known structure. These approaches estimates the model parameters of the network structure using statistical methods, and use these parameters to calculate the likelihood of the presence of a link between two nodes. Examples of probabilistic and statistical approaches include (Guimerà and Sales-Pardo 2009; Huang 2010; Wang, Satuluri, and Parthasarathy 2007). Algorithmic approaches directly uses the link-prediction as supervision to build the model. For example, link-prediction task can be formulated as a binary classification task where the positive instances are the pair of nodes which are connected in the network, and negative instances are the unconnected nodes. Examples include (Menon and Elkan 2011; Bliss et al. 2014). Unsupervised or self-supervised node embedding (such as DeepWalk (Perozzi, Al-Rfou, and Skiena 2014), node2vec (Grover and Leskovec 2016)), followed by training a binary classifier and Graph Neural network approaches such as GraphSage (Hamilton, Ying, and Leskovec 2017) belong to this category.

Distant-supervised credit-attribution

Various distant-supervised approaches have been developed for credit-attribution, but the prior have primarily focused on text documents. A document may be associated with multiple labels but all the labels do not apply with equal specificity to the individual parts of the documents. *Credit attribution* problem refers to identifying the specificity of labels to different parts of the document. Various probabilistic and neural-network based approaches have been developed to address the credit-attribution problem, such as Labeled Latent Dirichlet Allocation (LLDA) (Ramage et al. 2009), Partially Labeled Dirichlet Allocation (PLDA) (Ramage, Manning, and Dumais 2011), Multi-Label Topic Model (MLTM) (Soleimani and Miller 2017), Segmentation with Refinement (SEG-REFINE) (Manchanda and Karypis 2018), and Credit Attribution with Attention (CAWA) (Manchanda and Karypis 2020).

Another line of work uses distant-supervised credit-attribution for query-understanding in product search. Examples include, (i) using the reformulation logs as a source of distant-supervision to estimate a weight for each term in the query that indicates the importance of the term towards expressing the query's product intent (Manchanda, Sharma, and Karypis 2019a,b); and (ii) annotating individual terms in a query with the corresponding intended product characteristics, using the characteristics of the engaged products as a source of distant-supervision (Manchanda, Sharma, and Karypis 2020).

Citation-intent classification

There is a large body of work studying the intent of citations and devising categorization systems. In general, these approaches treat citation-intent classification as a text classification problem, and require the availability of training data with ground truth annotations. Representative examples include rule based approaches (Pham and Hoffmann 2003; Garzone and Mercer 2000) as well as machine-learning driven approaches (Valenzuela, Ha, and Etzioni 2015; Jurgens et al. 2018; Cohan et al. 2019). Generating labeled data for for these supervised approaches is difficult and time consuming, especially when the meaning of the labels is user-defined. In contrast, our approaches are distant supervised, that require no manual annotation.

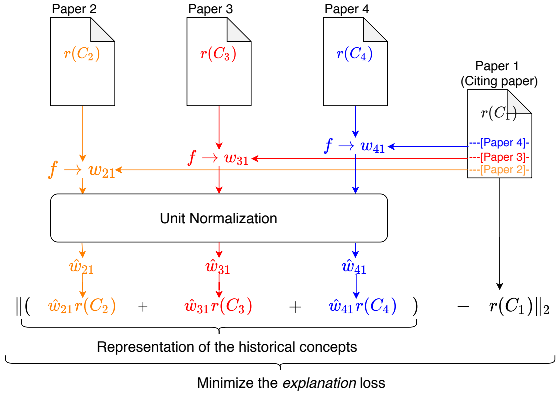

Figure 1: Overview of Content-Informed Index. Paper P_1 cites papers P_2, P_3 and P_4. The weights w_{21}, w_{31}, and w_{41} quantifies the extent to which P_2, P_3 and P_4 informs P_1, respectively. The function f is implemented as a Multilayer Perceptron.

3 Content-Informed Index (CII)

In the absence of labels that define the *impact*, we assume that the extent to which a cited paper informs the citing paper is an indication of the citation's impact. Specifically, we assume that each paper P_i can be represented as a set of *concepts* C_i. Further, we assume that each paper P_i is build on top of a set of historical concepts H_i, and its novelty N_i is the new set of concepts it proposes. The contribution of a cited paper P_j towards the citing paper P_i is the set of concepts $C_{ji} = C_j \cap H_i$. In other terms, the set of concepts C_i is given by:

$$C_i = N_i \cup H_i = N_i \cup [\cup_{P_i cites P_j} C_{ji}].$$

The task at hand is to quantify the extent to which C_{ji} contributes towards H_i. To achieve this task, we look into the following directions:

- **How do we supervise the exercise?** We minimize the novelty of paper P_i, by trying to explain the concepts in paper P_i (denoted by C_i) using the historical concepts, i.e., the concepts of the papers it cites (C_j). We call the loss associated with this minimization as the *explanation* loss. This gives rise to the following optimization problem:

$$\text{minimize} \sum_i N_i = \text{minimize} \sum_i C_i - H_i.$$

To proceed in this direction, we need to answer two questions, (i) How to represent the the set of concepts associated with the paper P_i?, and (ii) How do we represent the set of historical concepts H_i? As we show next, we use the textual content of the papers to estimate the representations of C_i and H_i. Thus, we formulate our problem as a distant-supervised problem, and the content of the papers acts as a source of distant-supervision.

- **How to represent the set of concepts associated with a paper?** For simplicity, we represent the set of concepts associated with a paper (C_i) as a pretrained vector representation (embedding) of its abstract, such as Word2Vec (Mikolov et al. 2013), GloVe (Pennington, Socher, and Manning 2014), BERT (Devlin et al. 2018), ELMo (Peters et al. 2018), etc. In this paper, we use the pretrained representations pretrained on scientific documents provided by ScispaCy (Neumann et al. 2019). The representation of C_i is denoted by $r(C_i)$.
- **How do we represent the set of historical concepts H_i?** As the set of historical concepts H_i is a union of the borrowed concepts from the cited papers (C_j), we simply represent the set of historical concepts as a weighted linear combination of the representation of the concepts of the cited papers, i.e.,

$$\begin{aligned} r(H_i) \quad &= \sum_{P_i \text{ cites } P_j} \tilde{w}_{ji} r(C_j) \\ \text{subject to} \quad &\sum_{P_i \text{ cites } P_j} \tilde{w}_{ji}^2 = 1 \\ &\tilde{w}_{ji} \geq 0; \forall (i,j). \end{aligned}$$

We have the constrained norm condition ($\sum\limits_{P_i \text{ cites } P_j} \tilde{w}_{ji}^2 = 1$) to make the representation of $r(H_i)$ agnostic to the number of cited-papers (a paper can cite multiple papers to reference the same borrowed concepts).

The weights $\tilde{w}_{ji}$ can be thought of as normalized similarity measure between the concepts of the cited paper, and the citation context. Thus, to estimate $\tilde{w}_{ji}$, we first estimate unnormalized $\tilde{w}_{ji}$, denoted by w_{ji}, and then normalize w_{ji} so as to have unit norm. The unnormalized weight w_{ji} is precisely the extent to which C_j contributes towards H_i (and hence C_i), i.e., the weight that we wish to estimate in this paper. We estimate w_{ji} as a multilayer perceptron, that takes as input the representations of the cited paper and the citation context. We use the representation associated with the corresponding concepts as the representations of the cited papers ($r(C_j)$). Similar to $r(C_j)$, we use the ScispaCy vector representation for the citation context as the representation of the context, and denote it by $r(j \to i)$.

The above discussion leads to the following formulation:

$$\begin{aligned} \underset{f}{\text{minimize}} \quad & \sum_i ||r(C_i) - \sum_{P_i \text{ cites } P_j} \tilde{w}_{ji} r(C_j)||^2 \\ \text{subject to} \quad & \tilde{w}_{ji} = \frac{w_{ji}}{\sqrt{\sum\limits_{P_i \text{ cites } P_j} w_{ji}^2}}; \forall (i,j), \\ & w_{ji} = f(r(C_j), r(C_{ji})); \forall (i,j), \\ & w_{ji} \geq 0; \forall (i,j), \\ & w_{ji} \leq b; \forall (i,j). \end{aligned} \tag{1}$$

The max-bound constraint ($w_{ji} \leq b$) is introduced to limit the projection space of the weights w_{ji}. This is because, without this constraint, for a given citing paper P_i, if the set of weights w_{ji} minimize Equation (1), then so will any scalar multiplication of the weights w_{ji}. This can potentially lead to the estimated weights being incomparable across different citing papers. Having a max bound on the estimated weights helps avoid this scenario. To take care of the constraints, the function $f(\cdot)$ can be implemented as a $L2-$regularized multilayer perceptron, with a single output node, and a non-negative mapping at the output node. Not that we do not explicitly set the max-bound b, but it is implicitly set by the $L2$ regularization of the weights of the function f. The $L2$ regularization parameter is treated as a hyperparameter. Figure 1 shows an overview of Content-Informed Index (CII).

4 Experimental methodology

Evaluation methodology and metrics

We need to evaluate how well the weights estimated by our proposed approach quantifies the extent to which a cited paper informs the citing paper. To this extent, we leverage various manually annotated datasets (explained later in Section 4), where the annotations quantify the extent of information in the citation. The task inherently becomes an ordinal association, and we need to evaluate how well the ranking imposed by our proposed method associates with the ranking imposed by the manual annotations. As a measure of rank correlation, we use the non-parametric Somers' Delta (Somers 1962) (denoted by Δ). Values of Δ range from $-1(100\%$ negative association, or perfect inversion) to $+1(100\%$ positive association, or perfect agreement). A value of zero indicates the absence of association. Formally, given a dependent variable (i.e., the predicted weights by our model) and an independent variable (i.e., the manually annotated ground truth), Δ is the difference between the number of concordant and discordant pairs, divided by the number of pairs with independent variable values in the pair being unequal.

Relation of Δ to other metrics: When the independent variable has only two distinct classes (binary variable), the area under the receiver operating characteristic curve (AUC ROC) statistic is equivalent to Δ (Newson 2002). Thus, Δ can also be visualized as a generalization of AUC ROC towards ordinal classification with multiple classes. Further, as the dependent variable (the weights estimated by our proposed approach) is real valued, having two tied values on the independent variable is very difficult. Thus, for our case, Δ is equivalent to Goodman and Kruskal's Gamma (Goodman and Kruskal 1959, 1963, 1972, 1979), and just a scaled variant of Kendall's τ coefficient (Kendall 1938), with are other popular measures of ordinal association.

Baselines

We choose representative baselines from diverse categories as discussed below:

Link-prediction approaches: The citation weights that we estimate in this paper can also looked from the link-prediction perspective, i.e., assigning a score to every citation (link) in the citation graph, the score portraying the likelihood of the existence of a link. Thus, the citations that are noisy, i.e., the edges that do not make sense with the respect to underlying link-prediction model get smaller weights. We compare against two link-prediction methods, one based on classic network embedding approach, and other belonging to recent Graph Neural Network (GNN) based approaches.

- DeepWalk (Perozzi, Al-Rfou, and Skiena 2014): DeepWalk is a popular method to learn node embeddings. DeepWalk borrows ideas from language modeling and incorporates them with network concepts. Its main proposition is that linked nodes tend to be similar and they should have similar embeddings as well. Once we have node embeddings as the output of DeepWalk, we train a binary classifier, with the positive instances as the pairs of nodes which are connected in the network, and negative instances are the unconnected nodes (generated using negative sampling). We provide results using two different classifiers: Logistic Regression (denoted by DeepWalk+LR) and Multilayer Perceptron (denoted by DeepWalk+MLP). Note that Deepwalk is a transductive model, and only considers the network topology, i.e., DeepWalk does not use the content of the papers to estimate the model.
- GraphSage (Hamilton, Ying, and Leskovec 2017): GraphSAGE is a Graph Concolutional Network (GCN) based framework for inductive representation learning on large graphs. GraphSage is trained with the link-prediction loss, so we do not use a second step (as in DeepWalk) to train separate classifier. Note that, GraphSage is an inductive model, so also considers the content of the papers in addition to topology of the network to estimate the model.

Text-similarity based baselines: We can think of the function f as a similarity measure between the cited paper and the citation context. Thus, we consider the following similarity measures as our baselines: We use the same pretrained representations as we used as an input to CII, and cosine similarity as the similarity measure, which is a popular similarity measure for text data.

- Similarity-Abstract-Context: Similarity between the cited abstract and the citation context.
- Similarity-Context-Abstract: Similarity between the citing abstract and the citation context.
- Similarity-Abstract-Abstract: Similarity between the cited abstract and citing abstract.

To calculate each of the above similarity measures, we use the same pretrained representations as we used as an input to CII, and cosine similarity as the similarity measure, which is a popular similarity measure for text data. The baselines belonging to this category can also be thought of as similarity-based link prediction approaches.

In addition, we also consider another simple baseline, referred to as *Reference Frequency*, where we assume that more frequently the cited paper is referenced in the citing paper, the higher the chances of the cited paper informing the citing paper. This assumption has also been used as a feature in prior supervised approaches (Valenzuela, Ha, and Etzioni 2015). The absolute frequency of referencing a cited-paper may provide a good signal regarding the information borrowed from the cited paper, when comparing with other papers being cited by the same citing paper. However, as the citation-behavior differs between papers, the absolute frequency may not be comparable across different citing papers. Thus, we also provide results after doing normalization of the absolute frequency of the citation references for each citing paper. We provide results for mean, max, and min normalization. Specifically, given a citation and the corresponding citing paper, the information weight for a citation is calculated by dividing the number of references of that citation, by the mean, max, and min of references of all the citations in that citing paper, respectively.

Datasets

The Semantic Scholar Open Research Corpus (S2ORC): The S2ORC (Lo et al. 2020) dataset is a citation graph of 81.1 million academic publications and 380.5 million citation edges. We only consider the publications for which full-text is available and abstract contains at least 50 words. This leaves us with a total of $5,653,297$ papers, and $30,533,111$ edges (citations).

ACL-2015: The ACL-2015 (Valenzuela, Ha, and Etzioni 2015) dataset contains 465 citations gathered from the ACL anthology[9], represented as tuples of (cited paper, citing paper), with ordinal labels ranging from 0 to 3, in increasing order of importance. The citations were annotated by one expert, followed by annotation by another expert on a subset of the dataset, to verify the inter-annotator agreement. We only use the citations for which we have the inter-annotator agreement, and the citations are present in the S2ORC dataset we described before. The selected dataset contains 300 citations among 316 unique publications. The total number of unique citing publications are 283 and the total number of unique cited publications are 38.

ACL-ARC: The ACL-ARC (Jurgens et al. 2018) is a dataset of citation intents based on a sample of papers from the ACL Anthology Reference Corpus (Bird et al. 2008) and includes 1,941 citation instances from 186 papers and is annotated by domain experts. The dataset provides ACL IDs for the papers in the ACL corpus, but does not provide an identifier to the papers outside the ACL corpus, making it difficult to map many citations to the S2ORC corpus. However, it provided the titles of those papers, and we used these titles to map these papers to the papers in the S2ORC dataset, if we found matching titles. The annotations in ACL-ARC are provided at individual citation-context level, leading to multiple annotations for some of the (cited paper, citing paper) pair. If this is the case, we chose the highest-informing annotation for such (cited paper, citing paper) pairs. The selected dataset contains 460 citations among 547 unique publications. The total number of unique

[9] https://www.aclweb.org/anthology/

Table 1: Results on the Somers' Δ metric.

Model	ACL-2015	ACL-ARC	SciCite
Content-Informed Index (CII)	$\mathbf{0.428 \pm 0.013}$	$\mathbf{0.308 \pm 0.010}$	$\mathbf{0.296 \pm 0.006}$
Ref. Frequency (Absolute)	0.325 ± 0.000	$\mathbf{0.308 \pm 0.000}$	0.144 ± 0.000
Ref. Frequency (Mean-normalized)	0.351 ± 0.000	0.300 ± 0.000	0.120 ± 0.000
Ref. Frequency (Min-normalized)	0.321 ± 0.000	0.298 ± 0.000	0.145 ± 0.000
Ref. Frequency (Max-normalized)	0.270 ± 0.000	0.172 ± 0.000	0.035 ± 0.000
Similarity-Abstract-Abstract	-0.041 ± 0.000	0.091 ± 0.000	-0.003 ± 0.000
Similarity-Abstract-Context	-0.147 ± 0.000	0.090 ± 0.000	-0.125 ± 0.000
Similarity-Context-Abstract	0.013 ± 0.000	-0.062 ± 0.000	-0.202 ± 0.000
Deepwalk+LR	-0.071 ± 0.016	0.190 ± 0.006	-0.037 ± 0.018
Deepwalk+MLP	-0.026 ± 0.011	0.205 ± 0.024	-0.047 ± 0.015
GraphSage	0.023 ± 0.045	0.132 ± 0.024	0.049 ± 0.019

citing publications are 145 and the total number of unique cited publications are 413.

SciCite (Cohan et al. 2019) SciCite is a dataset of citation intents based on a sample of papers from the Semantic Scholar corpus[10], consisting of papers in general computer science and medicine domains. Citation intent was labeled using crowdsourcing. The annotators were asked to identify the intent of a citation, and were directed to select among three citation intent options: Method, Result/Comparison and Background. This resulted in a total $9,159$ crowdsourced instances. We use the citations that are present in the S2ORC dataset we described before. Similar to ACL-ARC, the annotations are provided at individual citation-context level, leading to multiple annotations for some of the (cited paper, citing paper) pair. For such cases, we chose the highest-informing annotation for the (cited paper, citing paper) pairs. The selected dataset contains 352 citations among 704 unique publications. There is no repeated citing or cited publication in this dataset, thus, the total number of unique citing publications as well as unique citing publications are 352 each.

Parameter selection

We treat one of the evaluation datasets (ACL-ARC) as the validation set, and chose the hyperparameters of our approaches and baselines with respect to best performance on this dataset. For DeepWalk, we use the implementation provided here[11], with the default parameters, except the dimensionality of the estimated representations, which is set to 200 (for the sake of fairness, as the used 200 dimensional text representations for CII). For the models that require learning, i.e., the logistic regression part of Deepwalk, MLP part of Deepwalk, GraphSage, and CII, we used the ADAM (Kingma and Ba 2015) optimizer, with initial learning rate of 0.0001, and further use step learning rate scheduler, by exponentially decaying the learning rate by a factor of 0.2 every epoch. We use $L2$ regularization of 0.0001. The function f in CII was implemented as a multilayer perceptron, with three hidden layers, with 256, 64, and 8 neurons, respectively. We use the same network architecture for the MLP that we train on top of DeepWalk representations. We train the logistic regression and MLP parts of Deepwalk, GraphSage, and CII for a maximum of 50 epochs, and do early-stopping if the validation performance does not improve for 5 epochs. For GraphSage, we use the implementation provided by DGL[12]. We used mini-batch size of 1024 for training the models.

5 Results and discussion

Quantitative analysis

Table 1 shows the performance of the various approaches on the Somers' Delta (Δ) for each of the datasets ACL-2015, ACL-ARC and SciCite. For ACL-2015 and SciCite, the proposed approach CII outperforms the competing approaches; while for the ACL-ARC dataset, CII performs at par with the best performing approach. The improvement of CII over the second best performing approach is 22% and 103%, on the ACL-2015 and SciCite datasets, respectively.

Interestingly, the simplest baseline, Reference-frequency and its normalized forms are the second best performing approaches. While Reference-frequency performs at par with the CII on the ACL-ARC dataset, it does not perform as good on the other two datasets. This can be attributed to the fact that the number of unique citing papers in ACL-ARC dataset are relatively small. Thus, many citations in ACL-ARC are shared by the same citing paper, which is not the case with the other two datasets. Thus, as mentioned in Section 4, absolute frequency of referencing a cited-paper may provide a good signal regarding the information borrowed from the cited paper, when comparing with other papers being cited by the same citing paper. Further, even the normalized forms of the Reference-frequency lead to only marginal increase in performance for the ACL-2015 and SciCite datasets. Thus, the simple normalizations (such as mean, max and min normalization used in this paper), are not sufficient to address the difference in citation-behavior that occurs between different papers.

[10] https://www.semanticscholar.org/

[11] https://github.com/xgfs/deepwalk-c

[12] https://github.com/dmlc/dgl/blob/master/examples/pytorch/graphsage

Furthermore, we observe that simple similarity based approaches, such as cosine-similarity between pairs of various entities (each combination of citing abstract, citing abstract, and citation-context) performs close to random scoring (Δ value of close to zero). This validates that the simple similarity measures, like cosine similarity are not sufficient to manifest the the information that a cited-paper lends to the citing-paper; thus, showing the necessity of more expressive approaches, like CII.

In addition, the other learning-based link-prediction-based approaches perform considerably worse than the simple baseline reference-frequency. While on ACL-2015 and SciCite datasets, they perform close to random scoring, the performance on ACL-ARC dataset is better than the random baseline.

Qualitative analysis

In order to understand the patterns that the proposed approach CII learns, we look into the data instances with the highest and lowest predicted weights. As the function f takes as input both the abstract of the cited paper and the citation context, the learnt patterns can be a complex function of the cited paper abstract and the citation context. Thus, for simplicity, we limit the discussion in this section to understand the linguistic patterns in the citation context, and how these patterns associate with the weights predicted for them.

In this direction, we select $10,000$ citation-contexts corresponding to citations with highest predicted weights, and plot the word clouds for these contexts. We repeat the same exercise for the citation-contexts with the lowest predicted weights. Figures 2 and 3 shows the wordclouds for the highest weighted citations and lowest weighted citations, respectively. These figures show some clear discriminatory patters between the highest-weighted and lowest-weighted citations, that relate well with the information carried by a citation. For example, the words such as 'used' and 'using' are very frequent in the citation contexts of the highest weighted citations. This is expected, as such verbs provide a strong signal that the cited work was indeed employed by the citing paper, and hence the cited paper informed the citing work. Another interesting pattern in the highest weighted citations is the presence of words like 'fig', 'figure' and 'table'. Such words are usually present when the authors present or describe important concepts, such as methods and results. As such, citations in these important sections indicates that the cited work is used or extended in the citing paper, which signals importance.

On the other hand, the wordcloud for the least weighted citations (Figure 3) is dominated by weasle words such as 'may', 'many', 'however', etc. The words such as 'many' commonly occur in the related work section of the paper, where the paper presents some examples of other related works to emphasize the problem that the citing paper is solving. The words like 'may', 'however', 'but' etc are commonly used to describe some limitation of the cited work. Such citations are expected to be incidental, carrying less information, as compared to other citations.

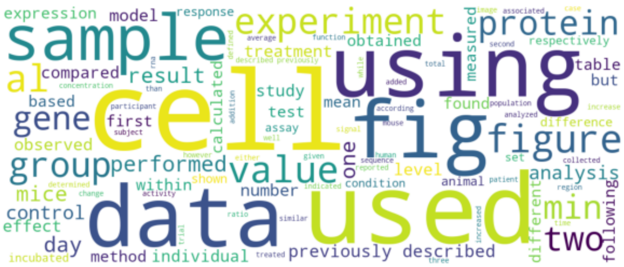

Figure 2: Word-cloud (Frequently occurring words) that appear in the citation context of the citations with the highest predict importance weights.

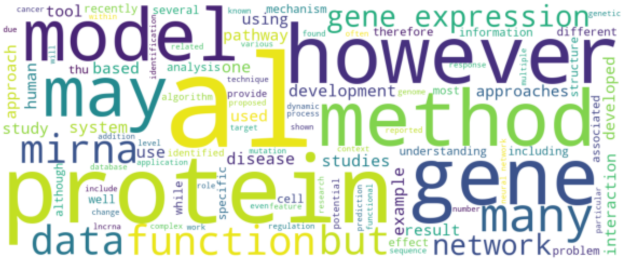

Figure 3: Word-cloud (Frequently occurring words) that appear in the citation context of the citations with the least predict importance weights.

6 Conclusion

In this paper, we presented approaches to estimate content-aware bibliometrics to accurately quantitatively measure the scholarly impact of a publication. Our distant-supervised approaches use the content of the publications to weight the edges of a citation network, where the weights quantify the extent to which the cited-publication informs the citing-publication. Experiments on the three manually annotated datasets show the advantage of using the proposed method on the competing approaches. Our work makes a step towards developing content-aware bibliometrics, and envision that the proposed method will serve as a motivation to develop other rigorous quality-related metrics.

References

Adamic, L. A.; and Adar, E. 2003. Friends and neighbors on the web. *Social networks* 25(3): 211–230.

Aguinis, H.; Suárez-González, I.; Lannelongue, G.; and Joo, H. 2012. Scholarly impact revisited. *Academy of Management Perspectives* 26(2): 105–132.

Bird, S.; Dale, R.; Dorr, B. J.; Gibson, B.; Joseph, M. T.; Kan, M.-Y.; Lee, D.; Powley, B.; Radev, D. R.; and Tan, Y. F. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics .

Bliss, C. A.; Frank, M. R.; Danforth, C. M.; and Dodds, P. S. 2014. An evolutionary algorithm approach to link prediction in dynamic social networks. *Journal of Computational Science* 5(5): 750–764.

Cohan, A.; Ammar, W.; van Zuylen, M.; and Cady, F. 2019. Structural Scaffolds for Citation Intent Classification in Scientific Publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3586–3596.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Ebesu, T.; and Fang, Y. 2017. Neural citation network for context-aware citation recommendation. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, 1093–1096.

Färber, M.; and Jatowt, A. 2020. Citation Recommendation: Approaches and Datasets. *arXiv preprint arXiv:2002.06961* .

Garzone, M.; and Mercer, R. E. 2000. Towards an automated citation classifier. In *Conference of the canadian society for computational studies of intelligence*, 337–346. Springer.

Giles, C. L.; Bollacker, K. D.; and Lawrence, S. 1998. CiteSeer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, 89–98.

Goodman, L. A.; and Kruskal, W. H. 1959. Measures of association for cross classifications. II: Further discussion and references. *Journal of the American Statistical Association* 54(285): 123–163.

Goodman, L. A.; and Kruskal, W. H. 1963. Measures of association for cross classifications III: Approximate sampling theory. *Journal of the American Statistical Association* 58(302): 310–364.

Goodman, L. A.; and Kruskal, W. H. 1972. Measures of association for cross classifications, IV: Simplification of asymptotic variances. *Journal of the American Statistical Association* 67(338): 415–421.

Goodman, L. A.; and Kruskal, W. H. 1979. Measures of association for cross classifications. In *Measures of association for cross classifications*, 2–34. Springer.

Grover, A.; and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864.

Guimerà, R.; and Sales-Pardo, M. 2009. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences* 106(52): 22073–22078.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in neural information processing systems*, 1024–1034.

Han, J.; Song, Y.; Zhao, W. X.; Shi, S.; and Zhang, H. 2018. hyperdoc2vec: Distributed representations of hypertext documents. *arXiv preprint arXiv:1805.03793* .

He, J.; Nie, J.-Y.; Lu, Y.; and Zhao, W. X. 2012. Position-aligned translation model for citation recommendation. In *International symposium on string processing and information retrieval*, 251–263. Springer.

He, Q.; Kifer, D.; Pei, J.; Mitra, P.; and Giles, C. L. 2011. Citation recommendation without author supervision. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 755–764.

He, Q.; Pei, J.; Kifer, D.; Mitra, P.; and Giles, L. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, 421–430.

Huang, W.; Kataria, S.; Caragea, C.; Mitra, P.; Giles, C. L.; and Rokach, L. 2012. Recommending citations: translating papers into references. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, 1910–1914.

Huang, W.; Wu, Z.; Liang, C.; Mitra, P.; and Giles, C. L. 2015. A neural probabilistic model for context based citation recommendation. In *Twenty-ninth AAAI conference on artificial intelligence*.

Huang, Z. 2010. Link prediction based on graph topology: The predictive value of generalized clustering coefficient. *Available at SSRN 1634014* .

Jurgens, D.; Kumar, S.; Hoover, R.; McFarland, D.; and Jurafsky, D. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics* 6: 391–406.

Kataria, S.; Mitra, P.; and Bhatia, S. 2010. Utilizing Context in Generative Bayesian Models for Linked Corpus. In *Aaai*, volume 10, 1. Citeseer.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika* 30(1/2): 81–93.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization URL http://arxiv.org/abs/1412.6980.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46(5): 604–632.

Kobayashi, Y.; Shimbo, M.; and Matsumoto, Y. 2018. Citation recommendation using distributed representation of discourse facets in scientific articles. In *Proceedings of the 18th ACM/IEEE on joint conference on digital libraries*, 243–251.

Li, H.; Councill, I.; Lee, W.-C.; and Giles, C. L. 2006. CiteSeerx: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web*, 883–884.

Liben-Nowell, D.; and Kleinberg, J. 2007. The link-prediction problem for social networks. *Journal of the American society for information science and technology* 58(7): 1019–1031.

LIU, Y.; YAN, R.; and YAN, H. 2016. Personalized Citation Recommendation Based on User's Preference and Language Model. *Journal of Chinese Information Processing* (2): 18.

Livne, A.; Gokuladas, V.; Teevan, J.; Dumais, S. T.; and Adar, E. 2014. CiteSight: supporting contextual citation recommendation using differential search. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 807–816.

Lo, K.; Wang, L. L.; Neumann, M.; Kinney, R.; and Weld, D. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–4983. Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

Manchanda, S.; and Karypis, G. 2018. Text segmentation on multilabel documents: A distant-supervised approach. In *2018 IEEE International Conference on Data Mining (ICDM)*, 1170–1175. IEEE.

Manchanda, S.; and Karypis, G. 2020. CAWA: An Attention-Network for Credit Attribution. In *AAAI*, 8472–8479.

Manchanda, S.; Sharma, M.; and Karypis, G. 2019a. Intent term selection and refinement in e-commerce queries. *arXiv preprint arXiv:1908.08564* .

Manchanda, S.; Sharma, M.; and Karypis, G. 2019b. Intent term weighting in e-commerce queries. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2345–2348.

Manchanda, S.; Sharma, M.; and Karypis, G. 2020. Distant-Supervised Slot-Filling for E-Commerce Queries. *arXiv preprint arXiv:2012.08134* .

Martínez, V.; Berzal, F.; and Cubero, J.-C. 2016. A survey of link prediction in complex networks. *ACM computing surveys (CSUR)* 49(4): 1–33.

Menon, A. K.; and Elkan, C. 2011. Link prediction via matrix factorization. In *Joint european conference on machine learning and knowledge discovery in databases*, 437–452. Springer.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Neumann, M.; King, D.; Beltagy, I.; and Ammar, W. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, 319–327. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/W19-5034. URL https://www.aclweb.org/anthology/W19-5034.

Newson, R. 2002. Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences. *The Stata Journal* 2(1): 45–64.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* .

Pham, S. B.; and Hoffmann, A. 2003. A new approach for scientific citation classification using cue phrases. In *Australasian Joint Conference on Artificial Intelligence*, 759–771. Springer.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 248–256. Association for Computational Linguistics.

Ramage, D.; Manning, C. D.; and Dumais, S. 2011. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 457–465. ACM.

Rokach, L.; Mitra, P.; Kataria, S.; Huang, W.; and Giles, L. 1978. A supervised learning method for context-aware citation recommendation in a large corpus. *INVITED SPEAKER: Analyzing the Performance of Top-K Retrieval Algorithms* 1978.

Soleimani, H.; and Miller, D. J. 2017. Semisupervised, multilabel, multi-instance learning for structured data. *Neural computation* 29(4): 1053–1102.

Somers, R. H. 1962. A new asymmetric measure of association for ordinal variables. *American sociological review* 799–811.

Tang, X.; Wan, X.; and Zhang, X. 2014. Cross-language context-aware citation recommendation in scientific articles. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 817–826.

Valenzuela, M.; Ha, V.; and Etzioni, O. 2015. Identifying meaningful citations. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*.

Wang, C.; Satuluri, V.; and Parthasarathy, S. 2007. Local probabilistic models for link prediction. In *Seventh IEEE international conference on data mining (ICDM 2007)*, 322–331. IEEE.

Yin, J.; and Li, X. 2017. Personalized citation recommendation via convolutional neural networks. In *Asia-Pacific web (APWeb) and web-age information management (WAIM) joint conference on web and big data*, 285–293. Springer.