

RESEARCH

Open Access



Leveraging citation intent for publication relatedness

Tuan Anh Phan¹ and Jason J. Jung^{1*}

*Correspondence:

Jason J. Jung

j2jung@gmail.com

¹Department of Computer Engineering, Chung-Ang University, 84 Heukseok-ro, Seoul 06974, South Korea

Abstract

This paper proposes a novel Intent-aware citation-based (IC) method for measuring publication relatedness using citation intents. Solely relying on citation counts and citation co-occurrence patterns, traditional citation-based methods treat all citations as equally informative and ignore citation intents, which reflects the rhetorical functions behind the citation contexts, leading to many tenuous publication links and obscuring the differential impact of citation intents. To address these limitations, our method discovers the citation intents for each citation context and retains meaningful inter-paper links only when they are derived from citation contexts sharing same intent. Our method is applied across three citation-based relations including Direct citation (IC_{DC}), Bibliographic coupling (IC_{BC}), Co-citation (IC_{CC}) and is evaluated through network clustering algorithms to assess textual coherence. Experimental results demonstrate that IC significantly outperforms traditional approaches while revealing the differential impact of citation intents. Particularly, the group of Comparison intents (*Differences* and *Similarities*) excels in most cases, with improvements of 0.0389 for IC_{BC} (NLP dataset) and 0.0659 for IC_{DC} (Biomedical dataset). Besides, *Extends*, *Motivation*, *Uses*, and *Future Work* also show strong potential, with *Extends* improving IC_{DC} by 0.0327 on the NLP dataset, and *Motivation* achieving a 0.0423 gain on the Biomedical dataset. Conversely, *Background* exhibits the worst performance among all intents, indicating minimal contribution to publication relatedness. Beyond effectiveness in paper–paper relationship modeling, our approach has broader implications for bibliometric analysis by enabling systematic investigations of how citation intent shapes scholarly relationships.

Keywords Publication relatedness, Citation intent, Citation network, Intent-aware citation network, Citation intent classification

1 Introduction

The exponential growth of scientific literature has increasingly complicated the task of measuring the relatedness between academic publications, which plays a crucial role in many areas of bibliometrics analysis, such as information retrieval [1–3], science mapping [4, 5], or research area clustering [6]. Measuring publication relatedness typically involves two main approaches: Citation methods, which leverage citation networks, and text-based methods, which rely on content similarity [7]. Aim to investigate the utility



© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

of citation information to identify related publications, our study concentrates on the citation-based approach.

Based on the citation network of articles, there are three traditional citation-based methods for publication relatedness such as Direct citation, Bibliographic coupling [8], and Co-citation [9] which are illustrated in Fig. 1. Direct citation (Fig. 1b) occurs when one publication explicitly references another, establishing straightforward connections. In contrast, Bibliographic coupling [8] and Co-citation [9] identify relatedness by examining pairs of publications that share common references or are cited together by other documents. In Fig. 1c, three papers *C*, *D* and *E* form a Bibliographic coupling network because papers *C* and *D* cite paper *A* while papers *D* and *E* cite paper *B*. Moreover, in Fig. 1d, a Co-citation network between paper *A* and *B* are formed as they are referenced together by three papers *C*, *D*, and *E*. Building on the core principles of Direct citation, Bibliographic coupling, and Co-citation, prior research has introduced diverse methodologies to quantify the strength of connections between publications [10–15]. To evaluate the effectiveness of large-scale methods, researchers construct citation networks using relatedness measures and cluster them, with the quality of resulting article clusters directly reflecting the accuracy and efficacy of publication relatedness measurement.

Although widely recognized within the academic community, the traditional citation-based approaches for calculating the publication relatedness generally assume that all citations contribute equally to scholarly relationships, without distinguishing the rhetorical functions they serve within citation contexts. Citation intent, defined synonymously with rhetorical functions [16, 17], refers to the author's purpose for referencing another work, such as providing background information, justifying research motivation, highlighting methodological similarities, contrasting findings, or suggesting future research directions. Conceptually, citation intent reflects the functional roles that cited papers play in the discourse of the citing work, offering a more fine-grained view of how scholarly publications are interconnected beyond mere citation presence.

By relying primarily on citation counts and citation co-occurrence patterns, traditional citation-based methods fail to distinguish the functional purposes of citation contexts and consequently treat all citations as uniformly informative signals, leading to two major limitations as follows:

- These traditional methods result in many weak or incorrect connections between articles. Indeed, without distinguishing different citation intents, papers that play different rhetorical roles with respect to the same citing or cited work may

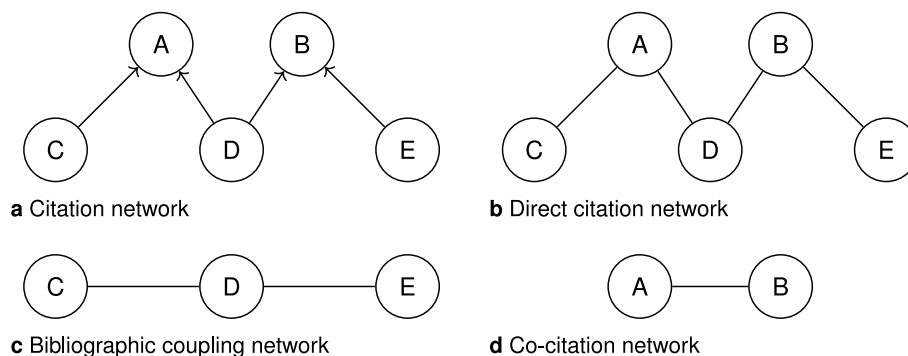


Fig. 1 Simple examples of: (a) citation network (b) Direct citation network (c) Bibliographic coupling network and (d) Co-citation network

nevertheless be linked together. Intuitively, papers that are connected to same paper through the same rhetorical function tend to exhibit higher content-related similarity, as they often engage with closely related concepts, methods, or research problems. In contrast, papers that link to a common work through different rhetorical functions are less likely to be closely related in terms of content or thematic focus, resulting in superficial and noisy links that degrade the quality of publication relatedness. For instance, when a work is cited with a *Background* intent, the cited content mainly serves a conceptual or contextual role, helping to frame the research problem or situate the study within prior literature. In contrast, when another work is cited by the same citing paper with a *Uses* intent, the cited content reflects a direct operational dependency, where specific methods, models, or experimental tools from the cited work are explicitly adopted in the citing paper. Therefore, although both works are cited by the same paper, the cited contents invoked in these two citation contexts are likely to differ substantially in their semantic and thematic focus, resulting in a weak link between these papers when modeled using traditional citation-based methods. This limitation leads to the intuition that one citation intent filtering method can enhance the semantic relatedness by better capturing shared functional roles underlying citation links.

- Traditional methods fall short in evaluating how citation intents specifically shape the strength of connections among scientific papers. Research, notably Valenzuela et al. [18], highlights that varying citation intents such as *Background*, *Comparison*, *Uses*, and *Extends* exert distinct influences on the relatedness of publications. In particular, intents like *Uses* and *Extends* tend to forge stronger links, with *Extends* notably reflecting significant methodological or conceptual advancements built upon prior work. Subsequent studies such as Hassan et al. [19–22] build on the work in Valenzuela et al. [18] to analyze the citation patterns. While these studies underscore the roles of citation intents in defining connection strength, they often depend on unverified assumptions and lack empirical validation. By treating all citation links as functionally equivalent, traditional citation-based methods are unable to capture the differing influences of citation intents on publication relatedness, thereby obscuring the nuanced roles that citations play in shaping scholarly connections.

To overcome the drawbacks of traditional citation-based methods, this study introduces an innovative Intent-aware citation-based (IC) approach to assess publication relatedness by leveraging the citation intents. First, the taxonomy for citation intents are defined based on the citation intent categories in Lauscher et al. [16]. After that, we develop a Citation intent recognition model to identify automatically the rhetorical intent from citation contexts. These predicted citation intents are then used to construct Citation intent graphs, which consider citation intent as the links between articles. Based on this Citation intent graph, we apply IC with three citation-based relations such as Direct citation (IC_{DC}), Bibliographic coupling (IC_{BC}), and Co-citation (IC_{CC}), compute Intent-aware citation connection scores, and construct Intent-aware citation networks. The core idea of IC is based on the intuition that, for a set of papers connected by citation-based links, papers that share the same citation intent, or equivalently, the same rhetorical function, are more likely to exhibit higher content relatedness, and are therefore retained. In contrast, links between papers associated with different citation intents tend to be weak and are thus removed. Particularly, for Direct citation, we differentiate

citation links based on their specific intents. For Bibliographic coupling, we connect papers that cite the same source papers only if they share the same citation intents, ensuring alignment in rhetorical purposes. Similarly, for Co-citation, we link papers cited together by a third paper only if they share the same rhetorical intent, avoiding superficial connections. Based on calculated scores, we formulate Intent-aware citation networks, which illustrate the relationship between articles with specific intents. Compared to traditional methods, our approach offers two key advantages. First, it captures publication connections based on the same citation intents, resulting in articles sharing rhetorical functions, leading to highly relevant and strongly related papers. Secondly, our approach allows for a precise evaluation of each citation intent's influence on publication relatedness by integrating distinct intents into citation networks, thereby enhancing the understanding of their specific contributions to scientific relationships.

To sum up, the main contributions of this paper are twofold as follows:

- We propose a novel Intent-aware citation-based approach to measure the publication relatedness. To the best of our knowledge, this is the first study to consider citation intent for this task.
- We evaluate our method on two datasets from the NLP and the Biomedical domains. The textual coherence results show that our method significantly outperforms traditional citation-based approaches and provides a solid quantitative assessment of the impact of each citation intent on measuring relatedness of academic publications.

The remainder of this paper is organized as follows. Section 2 reviews prior work on citation-based approaches for publication relatedness and the role of citation intent in identifying key citations. Section 3 presents the proposed method to build Intent-aware citation networks, detailing the Citation intent taxonomy, Citation intent classifier, and Intent-aware citation measures. Section 4 describes the experimental setup, covering datasets, evaluation metrics, and implementations. Section 5 reports the main results about Citation intent prediction and Textual coherence. Section 6 provides further analysis of the proposed approach. We conclude our paper and give some future directions in final section.

2 Literature review

2.1 Citation-based methods for measuring publication relatedness

Most citation-based studies on publication relatedness are grounded in three fundamental types of citation relations: DC, CC, and BC. Based on these foundational definitions, researchers primarily aimed to examine and compare the effectiveness of each relation, or to introduce modest adaptations to their formulations in order to suit their specific tasks or research objectives. Particularly, Boyack and Klavans [23] compared the accuracies of cluster solutions of a large corpus from biomedical literature using four similarity approaches: direct-citation, co-citation, bibliographic coupling, find out that direct citation yields the worst text coherence. Waltman et al. [7] proposed a new method for evaluating the accuracy of clustering solutions obtained using relatedness measures and finds that bibliographic coupling relations yield more accurate cluster solutions than direct citation and co-citation relations. Sjögårde and Ahlgren [13] evaluated five approaches to normalization of direct citation relations with respect to clustering solution quality. Their results clearly showed that normalization is preferred

over unnormalized direct citation relations and demonstrated that the fractional normalization approach, which can be considered the standard approach, causes inaccurate assignments. Sjögarde and Ahlgren [24] solved the seed-based information retrieval issue in network of research articles from NIH Open Citation Collection. The findings indicated that the Co-citation method surpasses both Bibliographic Coupling and Direct Citation relations in performance, with the optimal results achieved by integrating all Citation approaches. Wang and Sun [25] focused on metaphor comprehension, uses Bibliometrix to conduct a bibliometric study of research on metaphor comprehension from 2013 to 2023. The Co-citation analysis revealed that the Contemporary Theory of Metaphor (CTM) within Cognitive Linguistics serves as the primary intellectual cornerstone for studies on metaphor comprehension. Xie and Waltman [26] conducted a comparative analysis for science mapping with topic modeling and Citation clustering. In this study, a hierarchical classification system was built based on direct citation relation with the Leiden algorithm for network clustering. As a result, Co-citation demonstrated superior effectiveness in mapping the intellectual framework of cardiovascular diseases, effectively highlighting the presence of scientific micro-communities. Overall, these methods remain largely grounded in traditional citation-based relations and do not explicitly model citation intent at the level of individual links. As a result, papers that play different rhetorical roles with respect to the same citing or cited work may nevertheless be linked together, leading to numerous weak connections when modeling publication relatedness.

Another line of research extends traditional citation-based relations through the integration of textual information, where text-derived signals serve as auxiliary cues for refining citation-based connections. Particularly, Hamedani et al. [11] proposed a novel method that analyzed the directed citation to compute the contribution of cited papers to citing papers, and then combined this information with text-based information to compute the similarity of scientific papers. Zhang et al. [12] incorporated the frequency and distribution of documents cited in citing papers to compute the similarity matrix of the direct citation network, and combined it with the information from co-citation, bibliographic coupling, and text-based methods for scientific document clustering. Ahlgren et al. [27] analyzed the effect of enhancing direct citations, with respect to publication-publication relatedness measure by indirect citation relations and text relations on clustering solution and found that the extended direct approach had the best performance, followed by an approach in which direct citations are enhanced by the BM25 textual relatedness measure. Bascur et al. [15] used the extended direct citation from [27] as a citation metric for clustering the Biomedical publications from the PubMed database to analyze how different topics are represented in a science map. In these approaches, textual information served only as an auxiliary signal for adjusting link weights, while the links established between two papers under traditional citation-based methods remained unchanged. Consequently, these methods inherited the limitations of models that rely solely on traditional citation-based relations as above, which can result in many weak connections between publications due to the lack of homogeneity in the citation intents.

To address this issue, our approach leverages citation intent information to retain meaningful links and remove weak ones. Specifically, only connections between papers that built based on the citation contexts with same citation intents are preserved. This

filtering process helps eliminate noisy signals between papers and improves the quality of publication relatedness modeling.

2.2 Citation intent for identifying important citations

Several studies within the same line of research suggest that different citation intents are associated with varying strengths of linkage between scientific papers, as reflected through the distinction between important and unimportant citations. Valenzuela et al. [18] argued that important citations may indicate that the cited work is used or extended in the new effort and are more important than citations indicating related work or comparisons. Hassan et al. [19] extended the idea of Valenzuela et al. [18] by enhancing their features with six novel features (including context-based, cue words, and textual-based) to obtain improved performance with 0.91 AUC in overall classification accuracy. Hassan et al. [28] considered 450 citations from the Association for Computational Linguistics corpus, classified by experts as either important or unimportant, and extracted 64 features based on the methods of four techniques in Valenzuela et al. [18], Hassan et al. [19], Teufel et al. [29] Abu-Jbara et al. [30]. They applied the Extra-Trees classifier to select 29 best features and applied the Random Forest and Support Vector Machine classifiers to all selected techniques, obtaining an improvement by 11.25% in compared to baselines and 89% Precision-Recall area under the curve. Besides, they also proposed using an LSTM network, which uses 64 features to distinguish important and unimportant citations with 92.57% accuracy. Qayyum and Afza [20] presented a supervised learning method for important citations classification that is dominated by a set of distinct metadata features such as titles, authors, keywords, and references, and achieved an improved value of precision just by relying on freely available metadata. Aljohani et al. [31] employed the model to classify citations are important or unimportant with 15 novel features: contextual, quantitative, and qualitative, showing that the RF model outperforms the comparative models by 9.52%, achieving a 92% precision-recall area under the curve. An et al. [21] exploited the potential of combining both generative and discriminative models. Their study first used a topic model to generate the generative features, and then fed them to two discriminative traditional machine-learning models, SVM and RF (random forest), and a deep learning model, a convolutional neural network (CNN), with 13 other traditional features to identify important citations. James et al. [22] utilized a set of statistical, contextual, and metadata-based features to assess the influence of cited works on the current work, showing that sentiment analysis remarkably improves citation classification. Hannachi et al. [32] proposed a novel method by exploiting a set of features, including the similarity between citing and cited articles, and the similarity between the citation context and the abstract of the cited paper. The evaluation showed that their proposed model gives better results, especially with Random Forest, where the AUC-PR reached 0.90. Iqbal et al. [33] conducted a comparative analysis of various window sizes for in-text citation, aiming to select an optimal word embedding window size to identify crucial citations effectively.

While prior studies have suggested that citation intent may influence the strength of paper connections, such claims are often based on implicit assumptions rather than systematic evaluation [18]. In particular, citation intent has rarely been operationalized as an explicit signal for measuring publication relatedness, and its differential impact across intent types has not been empirically validated. This gap motivates an intent-aware

approach that directly incorporates citation intent into relatedness modeling and evaluates its effects through controlled experiments. To fill this gap, by intent-aware filtering of citation-based links, our method provides empirical evidence of how different citation intent influences and contribute unequally to the strength of publication relatedness.

3 Intent-aware citation-based method (IC)

In this section, we introduce our proposed method (IC) for publication relatedness. Specifically, we first present the Taxonomy of citation intent, which forms the foundation of our framework. Next, we describe the design of Intent-aware citation measures and explain how citation intent is integrated to construct the Intent-aware citation network. Finally, we detail the development of our Citation intent classifier, a deep learning-based model to automatically predict citation intents from citation contexts.

3.1 Taxonomy for citation intent

The foundation of our method relies on a robust definition of citation intent, which we adopt from the MultiCite dataset [16]. We chose the MultiCite dataset because it provides a rich taxonomy of citation intents and supports multi-label contexts, offering a more realistic basis for Intent-aware citation network analysis compared to earlier datasets. Particularly, the citation intent taxonomy and their corresponding examples are given as follows:

- *Background*: The target paper provides foundational knowledge, context, or overview relevant to the domain of the source paper.
- *Motivation*: The target paper highlights needs, goals, or research gaps that justify the problem statement or research direction of the source paper.
- *Uses*: The source paper directly employs ideas, methods, tools, datasets, or techniques from the target paper.
- *Extends*: The source paper builds upon, improves, or modifies concepts, methods, tools, or frameworks from the target paper.
- *Similarities*: The source paper draws parallels or notes commonalities between itself and the target paper, or between the target paper and other publications.
- *Differences*: The source paper emphasizes contrasts or distinguishes its contributions from the target paper or other related works.
- *Future Work*: The target paper proposes prospective research directions or speculative extensions that influence the source paper's outlook on potential studies.

In following parts, we will introduce how we use the citation intent to construct the Intent-aware citation network and how our network improves the traditional network in publication relatedness. To explain the our proposed method, we present a real example with articles extracted from Pubmed dataset, as shown in Table 1.

Based on the selected articles, a traditional citation network is built as Fig. 2.

Based on above defined citation intents and extracted papers, we extract the citation contexts and their corresponding citation intents, as shown in Table 2.

Given these citation intents, we build a Citation intent graph, where the nodes represent the papers, and the edges denote the citation intents in Fig. 3. Both traditional citation graph and our Citation intent graph are used to calculate traditional and Intent-aware relation scores for all article pairs.

Table 1 An example of articles extracted from Pubmed dataset

Paper	Reference	Title
P_1	Houghton et al. [34]	Elastin fragments drive disease progression in a murine model of...
P_2	Assi et al. [35]	Immune checkpoint inhibitors in advanced non–small cell lung...
P_3	Jackson et al. [36]	Analysis of lung tumor initiation and progression using conditional...
P_4	Tran et al. [37]	Immunogenicity of somatic mutations in human gastrointestinal...
P_5	Sade-Feldman et al. [38]	Defining T cell states associated with response to checkpoint...
P_6	Angermueller et al. [39]	Parallel single-cell sequencing links transcriptional and epigenetic...
P_7	Pelka et al. [40]	Spatially organized multicellular immune hubs in human...
P_8	Lieberman [41]	Granzyme A activates another way to die
P_9	Peng et al. [42]	Specific expression of GPR56 by human cytotoxic lymphocytes
P_{10}	Baixauli et al. [43]	The mitochondrial fission factor dynamin-related protein 1...
P_{11}	Wang et al. [44]	Differential U2AF1 mutation sites, burden and co-mutation genes...
P_{12}	Zheng et al. [45]	Transcriptomic profiles of neoantigen-reactive T cells in human...
P_{13}	Hanada et al. [46]	A phenotypic signature that identifies neoantigen-reactive T cells...
P_{14}	Marwitz [47]	Lung cancer subtypes generate unique immune responses.

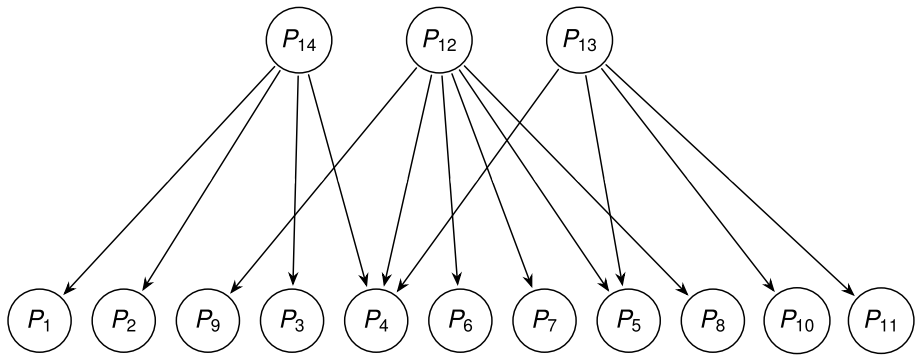


Fig. 2 Traditional citation graph created from selected articles

Table 2 Examples of citation contexts with their corresponding citation intents

Citation link (Citing → Cited)	Citation Context	Citation Intent
$P_{12} \rightarrow P_4$	We first determined whether neoantigen-..	Uses
$P_{12} \rightarrow P_4$	Using a conventional neoantigen-screening..	Similarities
$P_{12} \rightarrow P_5$	Single-cell sequencing has emerged as a..	Background
$P_{12} \rightarrow P_7$	Human melanoma (Li et al., 2019), human..	Uses
$P_{12} \rightarrow P_8$	GZMA encodes a serine protease found in..	Background
$P_{12} \rightarrow P_8$	GZMA encodes a serine protease found in..	Differences
$P_{12} \rightarrow P_9$	Notably, both exhausted neoantigen-..	Background
$P_{12} \rightarrow P_9$	In T cells, GPR56 has been reported to be..	Motivation
$P_{12} \rightarrow P_{15}$	To determine whether neoantigen-reactive..	Uses
$P_{13} \rightarrow P_4$	TCR gene-transduction was conducted as..	Uses
$P_{13} \rightarrow P_5$	A study on melanoma TILs reported that..	Background
$P_{13} \rightarrow P_{10}$	One out of the four genes that were down-..	Background
$P_{13} \rightarrow P_{11}$	Patient one had a CD4 reactivity to a mutant..	Background
$P_{14} \rightarrow P_1$	Lung tissue specimens were collected and..	Uses
$P_{14} \rightarrow P_2$	Newly developed immune checkpoint..	Background
$P_{14} \rightarrow P_3$	Kras ^{LSL-G12D} (i.e., Kras) (26), Trp53 ^{F1/F1} ..	Uses
$P_{14} \rightarrow P_4$	Furthermore, tumor-infiltrating lymphocyte..	Background

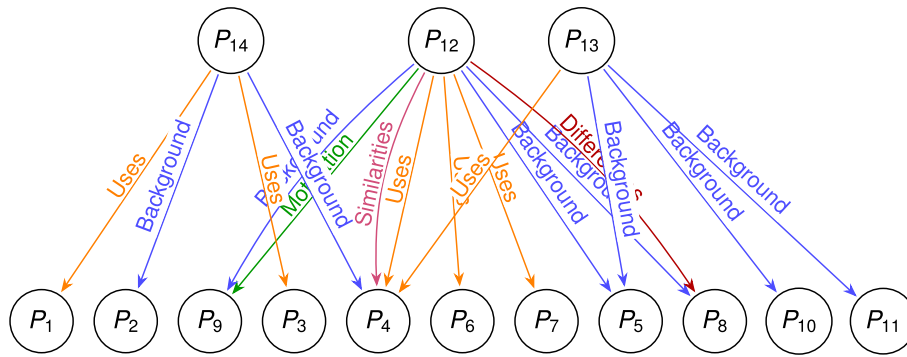


Fig. 3 Citation intent graph created from selected articles

3.2 Intent-aware citation measures

Building upon the Citation intent taxonomy and the Citation intent graph, our approach aim to build Intent-aware citation network by conditioning relatedness on specific intents, ensuring that connections are established only when intents align. To formalize, we denote $c_{ij} = 1$ as the presence of a citation from paper i th to paper j th and $c_{ij} = 0$ if paper i th does not cite paper j th. Similarly, we denote the presence of a citation from paper i th to paper j th with citation intent k th as $c_{ij}^k = 1$ (or $c_{ij}^k = 0$ for absence). Below, we detail each citation relation type, comparing the traditional approach with our proposed Intent-aware enhancement.

3.2.1 Direct citation

In conventional Direct citation analysis, relatedness is determined solely by the existence of a direct citation link between two papers. Formally, the relatedness score between papers i th and j th is: $r_{ij}^{DC} = \max(c_{ij}, c_{ji})$, where $r_{ij}^{DC} = 1$ if there are at least one paper cite another, and 0 otherwise. More specifically, in the above example, there are total 14 traditional Direct citation links are obtained. In contrast, our intent-aware Direct citation model refines this by incorporating the specific intent k th. The score is defined as: $(r_{ij}^{DC})^k = \max(c_{ij}^k, c_{ji}^k)$. Thus, $(r_{ij}^{DC})^k = 1$ if paper i th cites j th (or vice versa) with intent k th, and 0 otherwise. Our method establishes a connection only for matching intents, facilitating the assessment of each citation intent's influence on paper linkages. In the above example, there are 3 additional links are added because there are some pairs of papers connected with more than one citation intent (e.g., P_{12} and P_9 link together based on both *Background* intent and *Motivation* intent). Besides, in presented example, although P_{13} cites P_4 and P_{11} with different citation intents as *Uses* and *Background*, the traditional Direct citation still assigns same roles for both papers. However, P_{13} and P_4 are more closely aligned, as both former abstracts center on the identification and exploitation of neoantigen-reactive T-cell responses in solid tumors, specifically non-small cell lung cancer and gastrointestinal cancers, for advancing personalized immunotherapies, while P_{11} focuses on U2AF1 mutations and their prognostic implications in myelodysplastic syndrome without addressing immunological mechanisms.

3.2.2 Bibliographic coupling

Traditional Bibliographic coupling quantifies relatedness based on the overlap in reference lists, positing that shared references indicate thematic similarity. The score is typically normalized by the count of references:

$$r_{ij}^{BC} = \frac{\sum_{t=1}^{t=D} c_{it}c_{jt}}{\sum_{t=1}^{t=D} c_{it} + c_{jt} - c_{it}c_{jt}} \quad (1)$$

As a result, in the above example, there are 3 traditional Bibliographic coupling connections established from P_{12} , P_{13} , and P_{14} . In contrast, regarding Bibliographic coupling relation, we extend this model by restricting overlap to references sharing the same intent, thereby incorporating semantic nuance. The relatedness score between articles i th and j th with intent k th is defined as follow:

$$(r_{ij}^{BC})^k = \frac{\sum_{t=1}^{t=D} c_{it}^k c_{jt}^k}{\sum_{t=1}^{t=D} c_{it}^k + c_{jt}^k - c_{it}^k c_{jt}^k} \quad (2)$$

where the summation in both cases is computed over all publications in the database D , not only over the N publications that we are considering. Clearly, a connection exists only if intents match; otherwise, no relatedness is inferred, even for structurally shared references. In the above example, there are 2 connections removed because there are some cases that two papers cite another paper but different citation intents (e.g., P_{14} cites P_4 with *Background* intent and P_{12} also cites P_4 but with different intents such as *Similarities* and *Uses*, respectively). Besides, in our example, the traditional method creates a link between P_{14} and P_{12} due to sharing same reference P_4 . However, P_{14} and P_{12} are unrelated, as P_{14} investigates the host immune response to lung cancer in mouse models with specific mutational backgrounds, focusing on leukocyte content and PD-L1 expression, while P_{12} examines the transcriptomic profiles of neoantigen-reactive T cells in gastrointestinal cancers, emphasizing their exhausted state and distinct gene expression signatures. In contrast, due to different citation intents, our method does not assign any links between these articles.

3.2.3 Co-citation

Co-citation measures relatedness through the overlap in citing documents, assuming that papers cited together are intellectually linked. The standard score is calculated as:

$$r_{ij}^{CC} = \frac{\sum_{t=1}^{t=N} c_{ti}c_{tj}}{\sum_{t=1}^{t=N} c_{ti} + c_{tj} - c_{ti}c_{tj}} \quad (3)$$

As a result, there are total 26 traditional Co-citation connections established in our example. In contrast, our approach incorporates intent k th to filter shared citations, yielding:

$$(r_{ij}^{CC})^k = \frac{\sum_{t=1}^{t=N} c_{ti}^k c_{tj}^k}{\sum_{t=1}^{t=N} c_{ti}^k + c_{tj}^k - c_{ti}^k c_{tj}^k} \quad (4)$$

Compared to the conventional Co-citation method, two papers are considered as related only if they share the same citations with the same intents. In our example, our method

remove 15 connections because there are some papers that are cited by same another one but with different citation intents (e.g. P_7 and P_9 are both cited by P_{12} with *Uses* and *Background/Motivation* intents, respectively). Besides, in this example, according to the traditional Co-citation, P_7 and P_9 are strong related. However, in case of our method, they do not have any connections due to different citation intents. Our method's result is reasonable by in fact, P_7 and P_9 are unrelated as P_7 focuses on transcriptomic profiling and spatial organization of immune-malignant cell interactions in colorectal tumors, emphasizing MMRd and MMRp differences, while P_9 focuses on the expression and functional role of adhesion-GPCRs in human cytotoxic lymphocytes.

Based on the calculated Intent-aware scores, our method constructs Intent-aware citation networks, with scores indicating connection strength. For instance, we show our network with Co-citation relation for *Uses* and *Background* intents in Fig. 4a, b, respectively. We denote them as $IC_{CC-Uses}$ network and $IC_{CC-Background}$ network, respectively.

To assess our proposed model's ability to measure publication relatedness, we compare clustering outcomes between traditional citation networks and our method. Effective methods should group closely related articles, with cluster quality reflecting performance. Figure 5a, b display Co-citation results for traditional and proposed models, respectively, with gray boxes indicating article clusters.

Figure 5a show that by ignoring the citation intents, traditional method groups articles with different rhetorical functions into same cluster, leading to many weak or unrelated articles connections. For instance, P_7 with intent *Uses* and P_9 with intent *Background* nearly unrelated, as we mentioned in previous part. Grouping them into one cluster will harm the clustering quality, reduce the publication relatedness. In contrast, our method divides articles into separate groups with specific rhetorical functions, lead to more strong linkage within each cluster, enhance publication relatedness. More specifically, the Textual coherence, a metric for measuring clustering quality, detailed in Evaluation metrics subsection are computed for this example. We denote Coh_x for textual coherence of cluster consist of paper x . As a result, for traditional Co-citation method, $Coh_{4,5,10,11}$, $Coh_{6,7,8,9}$, and $Coh_{1,2,3}$ are 0.0187, 0.0197, and 0.0285, respectively. Meanwhile, for our IC_{CC} , $Coh_{5,10,11}$, $Coh_{8,9}$, $Coh_{2,4}$, $Coh_{4,6,7}$, and $Coh_{1,3}$ are 0.017, 0.0209, 0.0306, 0.0308, and 0.0364, respectively. This numbers indicate that our method yield clusters with higher quality than traditional method. Additionally, the textual coherence of clusters based on *Background* intent such as $Coh_{8,9}$, $Coh_{5,10,11}$, and $Coh_{2,4}$ are lower than that of *Uses* intent $Coh_{1,3}$ and $Coh_{4,6,7}$, also indicates that *Background* intent has a limitation in establishing the connections within articles compared to *Uses* intent.

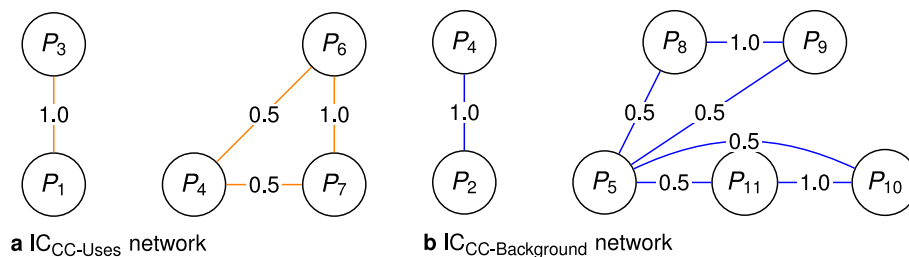


Fig. 4 Two Intent-aware Co-citation networks constructed from above example. **(a)** $IC_{CC-Uses}$ network using intent *Uses* versus **(b)** $IC_{CC-Background}$ network using intent *Background*

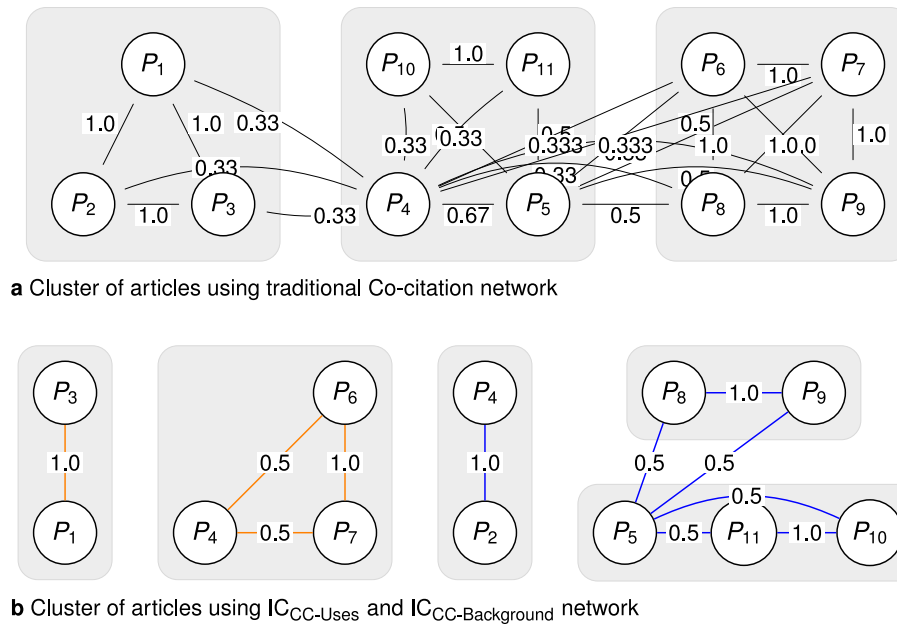


Fig. 5 Comparison of clusters of articles using (a) traditional Co-citation network and (b) our $IC_{CC-Uses}$ and $IC_{CC-Background}$ network. The gray boxes indicate article clusters

3.3 Citation intent classifier for scientific publication

To generate the intents category automatically for each citation sentence, we train a citation intent classification model for scientific articles, and use it to predict the corresponding citation intent labels. We use the MultiCite dataset [16] for the training dataset. To build the classifier, we adopt a pretrained language model as the backbone and fine-tune it in a supervised manner on MultiCite for citation intent classification. This model choice has been widely used and validated in prior studies on text classification, and has demonstrated strong and consistent performance. In particular, we employ SciBERT [48], a BERT-type pretrained language model to encode citation contexts, as its pretraining on large-scale scientific corpora has been shown to be well-suited for modeling domain-specific scientific text such as citation contexts. The overview of our citation intent classification model is presented in Figure 6.

The input citation contexts C are first split into a list of tokens $(TOK_1, TOK_2, \dots, TOK_{|C|})$ using SciBERT's tokenizer, where $|C|$ is the number of tokens in citation context. These tokens are concatenated with two special tokens $[CLS]$ and $[SEP]$, are then embedded as $(E_{[CLS]}, E_{TOK_1}, E_{TOK_2}, \dots, E_{TOK_{|C|}}, E_{[SEP]})$ thanks to SciBERT's Embedding layer. These vectors are fed through SciBERT's layers to obtain the contextual representation. SciBERT is a Transformer-based encoder model and encodes citation contexts using stacked self-attention layers, where the output of each layer is fed as input to the subsequent layer. This mechanism allows each token to attend to the entire input sequence, capturing the long-range contextual dependencies in scientific text and producing contextualized representations of citation contexts. Therefore, unlike traditional approaches that rely on manual feature engineering, feature selection in our model is implicitly performed during fine-tuning, as SciBERT learns task-specific representations directly from the input citation text. The representations for whole tokens on the top layer of SciBERT are denoted by $(h_{[CLS]}, h_{TOK_1}, h_{TOK_2}, \dots, h_{TOK_{|C|}}, h_{[SEP]})$. The embedding of the $[CLS]$ token

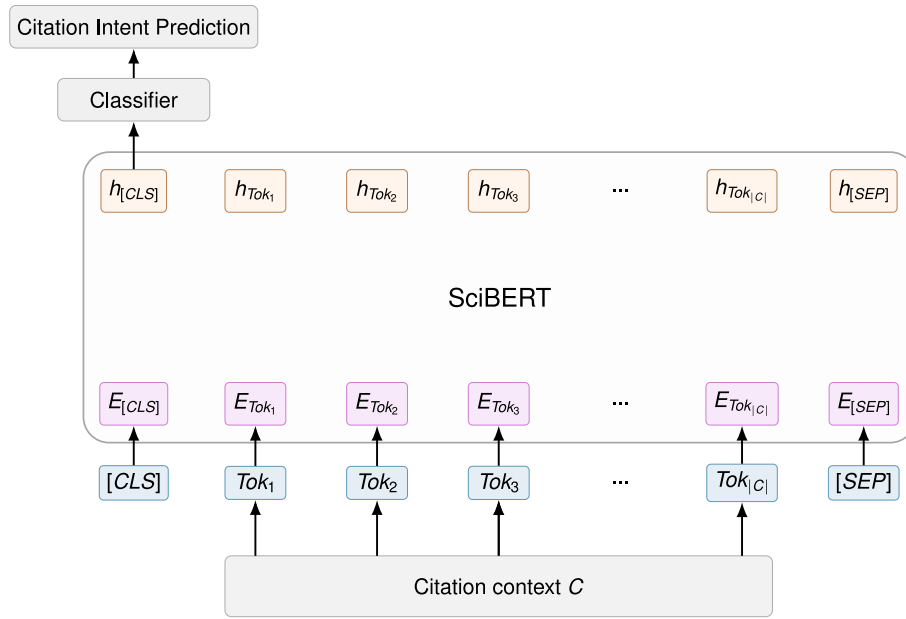


Fig. 6 Overview of our citation intent recognition model for scientific publication

$h_{[CLS]}$, is chosen for the representation of the whole citation context h_C and is used for classification. Because MultiCite is a multilabel dataset where each citation context can have more than one label, we use binary cross-entropy loss ($\mathcal{L}_{BCE}$) for the objective function. The representation of the whole citation context h_C and the binary cross-entropy loss is computed as follows:

$$h_C = h_{[CLS]} \in R^d \quad (5)$$

$$r_i = W^C h_C \in R^L \quad (6)$$

$$\hat{y}_i = \text{sigmoid}(r_i) = \left(\frac{1}{1 + e^{r_{ij}}} \right)_j \in (0, 1)^L \quad (7)$$

$$\mathcal{L}_{BCE} = - \sum_{i=1}^M \sum_{j=1}^L y_{ij} \log(\hat{y}_{ij}) \quad (8)$$

where R^d and L are the number of dimensions of the document representation vector and the number of intent categories, M is the number of samples in the training dataset, $W^C \in R^{L \times d}$ is a learnable parameter. Besides, $y_i \in \{0, 1\}^L$ and $\hat{y}_i \in (0, 1)^L$ are the gold binary labels and predicted probabilities of sample i th, where $y_{ij} = 1$ if intent j th is one of the gold labels of sample i th.

We train our model in a supervised manner on the MultiCite dataset. Training is conducted by minimizing the binary cross-entropy loss between predicted labels and gold annotations using the Adam optimizer [49]. Optimization is performed with mini-batch gradient descent for a fixed number of epochs. We employed grid search as a hyperparameter tuning strategy to identify the optimal values for the model hyperparameters based on performance on the development set. The classifier and hyperparameter configuration that achieve the highest strict metric (which is described in Citation intent

classification performance) on the development set are selected as the final model and hyperparameter values for evaluation on the test set. During inference, the model generates probabilities $\hat{y}_i$ for each citation intent, and a threshold of 0.5 is applied to determine the final multi-intent predictions. We choose this threshold because it is widely used and validated in prior studies on multi-label classification.

4 Experiments

4.1 Datasets

To train our citation intent classification model, we adopted the MultiCite dataset [16]. To conduct the experiments with the citation network, we selected two datasets from two academic disciplines, such as the Natural Language Processing domain (NLP) [50] and the Biomedical domains [12]. These were chosen for their widespread use in research, public availability, and access to full-text articles. Detailed information about each above datasets is given as follows.

4.1.1 MultiCite

The MultiCite dataset [16] is a dataset that consists of 12.6K citation contexts from 1.2K computational linguistics papers. Contrary to the traditional dataset for citation analysis, MultiCite provides the variable-length citation context (multi-sentence context) and multiple intents (multiple labels) for one citation context. The definition of citation intents in MultiCite exists in 7 types: *Background*, *Motivation*, *Uses*, *Extends*, *Similarities*, *Differences*, and *Future work*. For the citation intent classification task, the author split the dataset into 5491 samples for training, 2447 samples for validation, and 3313 samples for testing.

4.1.2 NLP dataset

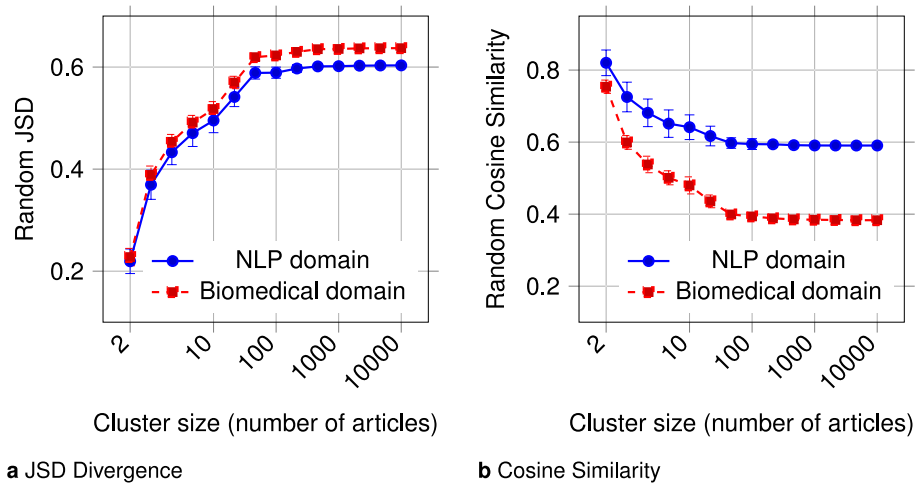
For the computational linguistics publications, we collect the dataset from the ACL Anthology corpus [50]. This dataset contains papers and posters from the ACL Anthology corpus, including complete PDFs, BibTeX files with abstracts, GROBID extraction results with full text and references, and a dataframe with metadata and full text for analysis. To conduct the experiments, we collect all of publications released from 2020. After removing the duplications, we have 15,455 articles and a total of 404,745 citation contexts.

4.1.3 Biomedical dataset

Following the work in Zhang et al. [12], we re-use their dataset, which consists of open-access and full-text scientific research articles in the field of oncology collected from 14 journals between the period 2007–2016 in the PubMed database (<https://www.ncbi.nlm.nih.gov/pmc>) and the PMC database (<https://www.ncbi.nlm.nih.gov/pubmed>). Documents whose article type includes letter, overview, or editorial are excluded, because the formats of letter and editorial are flexible and difficult to process. Totally, we collect 10,994 publications with 378,021 citation contexts. Detailed information about the number of documents by journal in the Biomedical dataset is given in Table 3.

Table 3 Number of research documents by journal in the Biomedical dataset

Journal title	Number of documents
Oncotarget	7457
Molecular Cancer	1340
Oncogene	1055
Leukemia	376
Oncolmunology	194
Annals of Oncology	155
Journal of the National Cancer Institute	145
Clinical Epigenetics	83
Breast Cancer Research	45
Neuro-Oncology	38
Journal of Thoracic Oncology	36
Cancer Cell	33
Stem Cells	22
Molecular Oncology	15
Total	10,994

**Fig. 7** JSD Divergence (a) and Cosine Similarity (b) of random sets of documents by cluster size in two datasets

4.2 Evaluation metrics

To evaluate the performance of the cluster solution, we follow within-textual coherence as the method in Boyack and Klavans [23]. We adopt this metric because it is widely used in this field and provides an accurate evaluation of semantic consistency within clusters, which aligns with the goals of our study. Specifically, in their study, textual coherence is measured using the Jensen-Shannon divergence [51], which calculates the distance between two probability distributions. Given the word distribution of document p and its cluster q , the JSD score is calculated as follow:

$$JSD(p, q) = 0.5 \times D_{KL}(p, m) + 0.5 \times D_{KL}(q, m) \quad (9)$$

where $m = (p + q)/2$, D_{KL} and $D_{KL}(p, m) = \sum p_i \log(p_i/m_i)$ is the Kullback-Leiler divergence. JSD is computed for each cluster as the average JSD over all documents in the cluster. Because JSD varies with cluster size, larger clusters will naturally be more divergent than smaller clusters. The maximum divergence cluster can be approximated by creating random clusters of documents from the corpus. Figure 7a shows the

computed JSD value and the variance as the error bar for randomly formed clusters in various sizes in both the NLP domain and the biomedical domain. In our study, we set the number of random clusters drawn from the corpus is 100. The small value in the error bars in two domains indicates that JSD is not a noisy measure. Based on the measured values, a curve fit is used to estimate the JSD value for every cluster size.

Coherence of cluster i th with cluster size of n_i is calculated as:

$$Coh_i = JSD_{rand}(n_i) - JSD_{actual} \quad (10)$$

where $JSD_{actual}(i)$ is the JSD divergence computed from the cluster i th in cluster solution, and $JSD_{rand}(n_i)$ is the JSD divergence value estimated from the random clusters. For the entire cluster solution, the average coherence value is calculated as weighted coherence over all clusters:

$$Coh = \frac{\sum_{i=1}^{N_C} n_i Coh_i}{\sum_{k=1}^{N_C} n_k} \quad (11)$$

where N_C is the number of clusters and n_i is the number of documents in cluster i th. Besides using JSD divergence, in this study, we also propose a new way to define the text coherence of a cluster: Coherence based on text Cosine Similarity. Particularly, we compute the cosine similarity between the vector representation of each document d with the vector representation of the cluster:

$$Sim(d, c) = \frac{\sum_i (d_i c_i)}{\sqrt{\sum_i d_i} \sqrt{\sum_i c_i}} \quad (12)$$

Cosine similarity of a cluster will be also calculated as the average score over all documents, then the text coherence of the cluster is the difference between the score of the random and actual cluster: $Coh_i = Sim_{actual}(i) - Sim_{rand}(n_i)$ where $Sim_{actual}(i)$ denote the Cosine similarity of cluster i th and $Sim_{rand}(n_i)$ is the Cosine similarity estimated from the random clusters from our corpus. Similar to JSD Divergence, the Cosine similarity is also calculated for randomly formed clusters with different cluster sizes and is shown in Figure 7b.

Text coherence score of the entire cluster solution is computed as follows:

$$Coh = \frac{\sum_{i=1}^{N_C} n_i Coh_i}{\sum_{k=1}^{N_C} n_k} \quad (13)$$

4.3 Implementation details

Regarding the citation intent classification model, we use the scibert-scivocab-uncased [48] with 12 layers, 768 hidden size, and 31,116 vocabulary size. For training setting, we use the Adam optimizer [49], with the learning rate is varied in $\{5e^{-6}, 1e^{-5}, 2e^{-5}\}$. We train our model using various batch size in $\{2, 4, 8, 16, 32\}$ with the maximum number of epochs 5.

After constructing both traditional and Intent-aware citation network, we use cluster algorithm to derive article clusters. In this study, we employ a well-known community detection method, proposed by Leicht and Newman [52], which optimizes the modularity function by maximising the difference between the actual number of edges in a

Table 4 Comparison of strict and weak evaluation on Dev and Test sets in the MultiCite dataset

Methods	Dev set		Test set	
	Strict	Weak	Strict	Weak
Lauscher et al. [16]	—	—	0.66	0.78
Our model	0.62	0.77	0.66	0.79

Table 5 Classification performance (Precision, Recall, and F1-score) on test set and number of train samples for each citation intent

Citation intent	Precision	Recall	F1	Train samples
Background	0.8721	0.7956	0.8321	2144
Uses	0.7737	0.7785	0.7761	1848
Extends	0.6319	0.4682	0.5379	365
Motivation	0.5274	0.5000	0.5133	433
Differences	0.6946	0.7083	0.7014	836
Similarities	0.6732	0.4296	0.5245	609
Future work	0.6667	0.4348	0.5263	37

community and the expected number of such edges. To optimize the objective function, we implement two well-known and are widely recognized in the academic community: the Louvain algorithm [53] and Leiden algorithm [54]. We conduct the cluster algorithm in all citation relations. For each case, we run 10 times with different random seeds, and the final results are reported as the average values.

5 Results

5.1 Citation intent classification performance

Following the work in Zhang et al. [16], we evaluate our Citation intent classification model using two metrics: the *strict version* (correct only if all predicted labels match gold labels), and the *weak version* (correct if at least one predicted label matches gold labels). The results in the Dev set and the Test set are given in Table 4.

Particularly, our model achieves a 0.62 score (strict version) and a 0.77 score (weak version) on the Dev set, while in the Test set, our model obtains comparable results with the work in Lauscher et al. [16] with 0.66 score and 0.79 score for strict and weak versions, respectively.

Table 5 reports the classification performance in terms of Precision, Recall, and F1-score together with the number of training samples for each citation intent. The results in the table indicate that *Background*, *Uses*, and *Differences* intents achieve the highest performance, primarily due to the abundance of training samples, which enables the model to learn stable and consistent representations for these intent categories. In contrast, intent types with fewer training instances, such as *Extends*, *Motivation*, and *Similarities*, exhibit lower performance. Notably, despite the very limited number of training samples available for *Future work*, the model still attains relatively competitive performance, suggesting that the proposed system is less sensitive to data imbalance even for rare citation intents.

The close alignment in the overall performance and the stable results across citation intent categories confirms the accuracy and robustness of our citation intent classifier, making it a reliable component for our system. Therefore, throughout this study, we use this model to predict citation intents.

Table 6 shows the distribution of Intent-aware connections, which are predicted across three citation relations: DC (Direct citation), CC (Co-citation), and BC (Bibliographic coupling) in two datasets.

It is clear that *Background* intent predominates in Bibliographic coupling, with over 2.9 million connections in NLP and 3.6 million in Biomedical, followed closely by the *Uses* intent (2.88 million for NLP, 1.07 million for Biomedical). Moreover, Bibliographic coupling produces far more connections than Direct citation and Co-citation in both domains. This phenomenon is reasonable because when calculating the relatedness of articles based on Direct citation and Co-citation relations, we only consider the citing papers and the cited papers in the considered dataset. The case for Direct citation is obvious because we aim to establish the connections between two articles in our corpus. For the case of Co-citation, we can not take into account the citing papers that do not provide the full-text of citation contexts. In contrast, Bibliographic coupling leverages full-text contexts from citing papers in our dataset, enabling inclusion of cited papers beyond the corpus.

In addition, Table 6 shows that there are several cases in which the number of relations with specific intents are quite small, which can affect the cluster results. Therefore, for each dataset, we do not compare the results with the intents that have the average number of relations per publication (ANR) lower than 0.01.

5.2 Textual coherence

In this section, we present textual coherence results for three citation relations across two datasets, measured using JSD and Cosine Similarity metrics with Louvain and Leiden clustering algorithms. Specifically, JSD_Louvain (blue) denotes the coherence score using the JSD metric with Louvain clustering, while JSD_Leiden, Cosine Similarity_Louvain, and Cosine Similarity_Leiden reflect corresponding scores for their respective metrics and algorithms. To facilitate comparison and result analysis, we introduce the term “group of Comparison intents” to refer to the experimental grouping of the *Similarities* and *Differences* intents, while preserving them as distinct categories in the formal taxonomy.

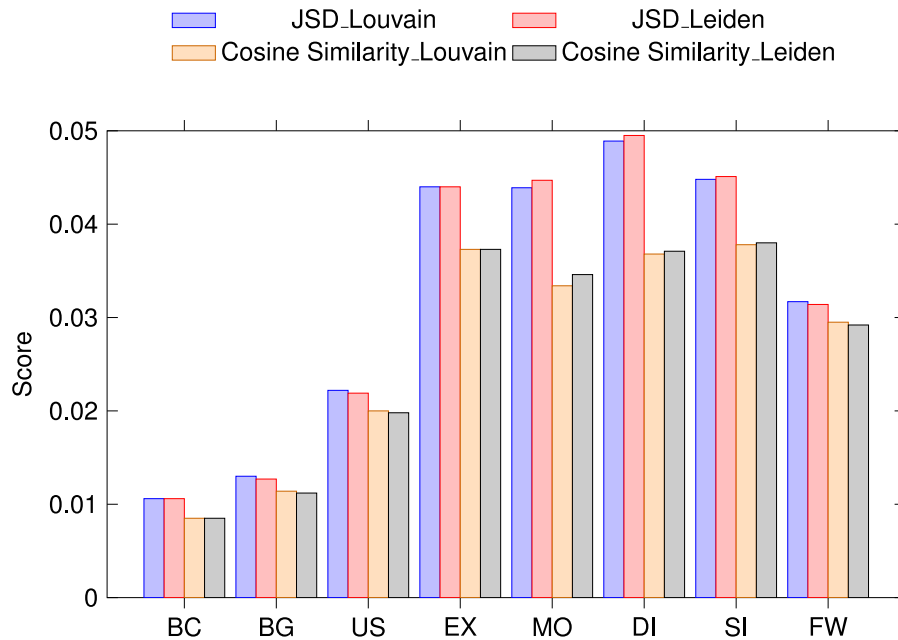
5.2.1 Bibliographic coupling

Figure 8 presents the textual coherence of Bibliographic coupling relations in the NLP domain. Clearly, the traditional Bibliographic coupling scores are relatively low across all settings. In contrast, our IC_{BC} enhances clustering quality, with the most significant

Table 6 Distribution of the number of intent-aware connections in three citation relations across the NLP dataset and Biomedical dataset

Citation Intents	NLP dataset			Biomedical dataset		
	DC	BC	CC	DC	BC	CC
Background	46,262	29,268,040	77,308	9920	3,660,574	5206
Uses	17,338	28,801,032	10,026	2122	1,079,524	308
Extends	1294	92,618	172	28	74	0
Motivation	4374	177,210	1828	730	40,726	56
Differences	3544	206,674	804	270	3390	10
Similarities	3980	228,332	1152	880	20,652	54
Future work	854	4538	222	110	220	14

Denotations as DC, BC, and CC stand for Direct citation, Bibliographic coupling, and Co-citation, respectively

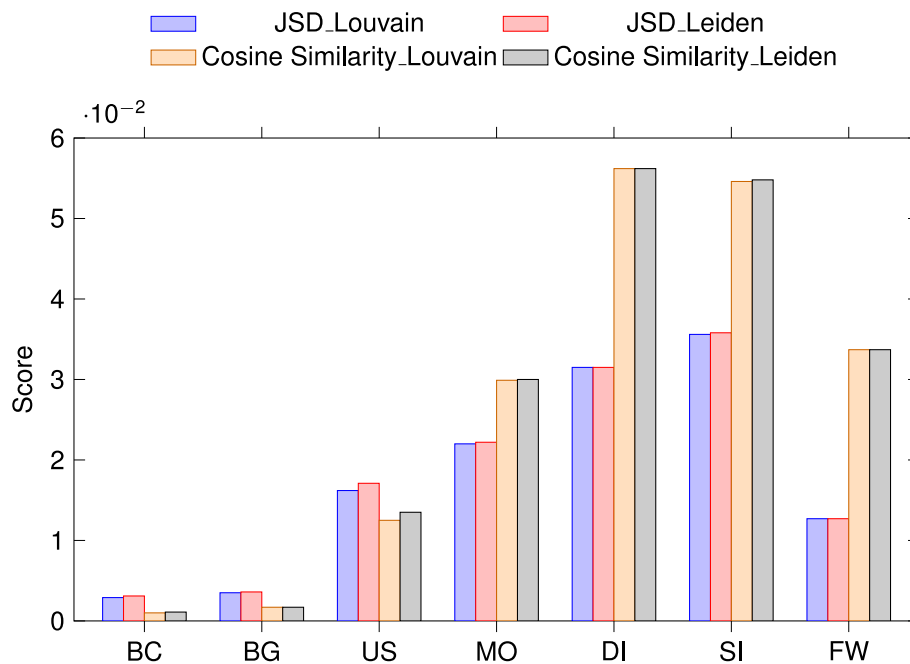


Codes: BC = traditional Bibliographic coupling;
 {BG, US, MO, DI, SI, FW} corresponds to IC_{BC} with citation intents {*Background*,
Uses, *Motivation*, *Differences*, *Similarities*, *Future work*}, respectively.

Fig. 8 Comparison of Bibliographic coupling methods in the NLP domain. **Codes:** BC = traditional Bibliographic coupling; {BG, US, MO, DI, SI, FW} corresponds to IC_{BC} with citation intents {*Background*, *Uses*, *Motivation*, *Differences*, *Similarities*, *Future work*}, respectively

improvements in coherence using JSD measure, observed with the group of Comparison intents, particularly $IC_{BC-Differences}$ (0.0489 for Louvain, 0.0495 for Leiden), obtains the enhancement by 0.0383 and 0.0389 compared to baseline model. For the textual coherence with Cosine Similarity measure, $IC_{BC-Extends}$ (0.0373 for both Louvain and Leiden) and $IC_{BC-Similarities}$ (0.0378 for Louvain, 0.0380 for Leiden) show the highest gains, with $IC_{BC-Differences}$ also performing competitively. The group of *Uses*, *Motivation*, and *Future work* shows moderate improvements, with textual coherence scores of approximately 0.022 with JSD for $IC_{BC-Uses}$, 0.0439 with JSD for $IC_{BC-Motivation}$, and 0.031 with JSD for $IC_{BC-Future\ work}$. For the *Background* intent, textual coherences are only slightly higher than the baseline values but smaller than those achieved by most other intents, suggesting it contributes the smallest discriminative power compared to other Intent-aware methods.

Figure 9 presents the performance of Bibliographic coupling methods in the Bio-medical domain, excluding the result for *Extends* intent because its ANR is only 0.0034, lower than the threshold 0.01. Similar to the NLP domain, the traditional Bibliographic coupling consistently yields the lowest performance across all settings, confirming its limited ability to capture meaningful publication relationships. Besides, all intent-incorporated variants outperform the baseline, with *Similarities* and *Differences* achieving the highest textual coherence: $IC_{BC-Similarities}$ tops JSD (0.0356 for Louvain, 0.0358 for Leiden), while $IC_{BC-Differences}$ lead in Cosine Similarity measure (0.0562 for both Louvain and Leiden). Moreover, *Uses*, *Future work*, and *Motivation* also show significant



Codes: BC = traditional Bibliographic coupling; {BG, US, MO, DI, SI, FW} corresponds to IC_{BC} with intents {*Background*, *Uses*, *Motivation*, *Differences*, *Similarities*, *Future work*}, respectively.

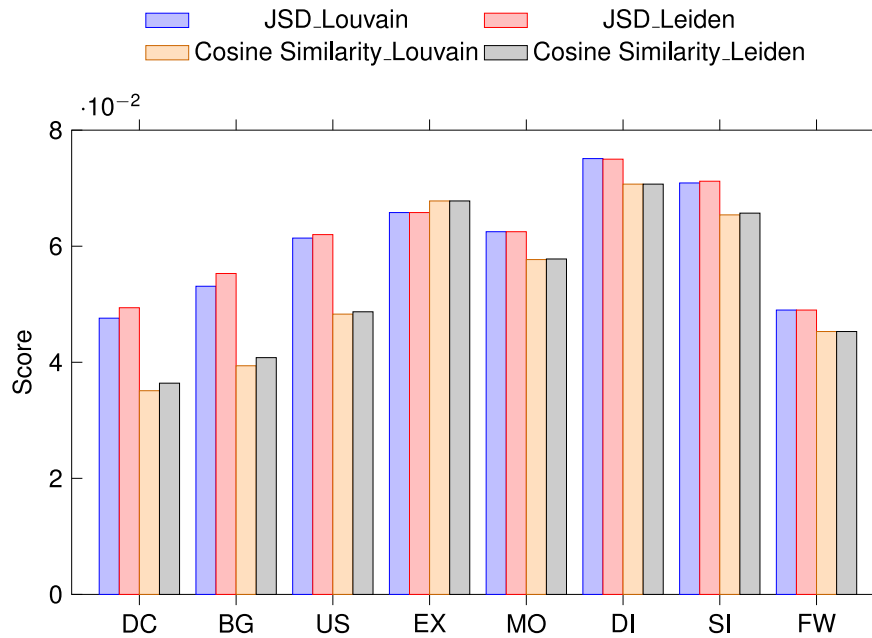
Fig. 9 Comparison of Bibliographic Coupling methods in the Biomedical domain. **Codes:** BC = traditional Bibliographic coupling; {BG, US, MO, DI, SI, FW} corresponds to IC_{BC} with intents {*Background*, *Uses*, *Motivation*, *Differences*, *Similarities*, *Future work*}, respectively

improvements, whereas using *Background* intent provides the weakest gains, reinforcing its limited effectiveness in enhancing Bibliographic coupling relations.

5.2.2 Direct citation

Figure 10 presents a comparison of Direct citation methods in the NLP domain. Overall, the traditional Direct citation method exhibits the poorest textual coherence, highlighting its limited ability to assess publication relatedness. The group of Comparison intents, including *Differences* and *Similarities*, significantly outperforms other methods, with $IC_{DC-Differences}$ intent achieving the highest coherence score with Cosine Similarity (0.0707 for both Louvain and Leiden) and JSD (0.0751 for Louvain, 0.0750 for Leiden), while $IC_{DC-Similarities}$ obtains coherence scores 0.0712 (JSD, Leiden) and 0.0657 (Cosine Similarity, Leiden). Using *Extends* intent also improves significantly, with Cosine Similarity-based coherence at 0.0678, obtains enhancement by 0.0327 coherence score. The other intents, such as *Motivation*, *Uses*, and *Future work* show moderate gains while using *Background* intent also shows a minimal improvement over the baseline, yet it ranks as the lowest-performing intent compared to other Intent-aware methods.

Figure 11 presents a detailed comparison of Direct citation on the Biomedical domain, excluding *Extends*, *Differences*, and *Future work* due to ANRs below the 0.01 threshold. Consistent with the NLP dataset, the traditional method performs weakest, with a textual coherence score using JSD of 0.0370 for Louvain and 0.0373 for Leiden, and Cosine Similarity of 0.0365 for Louvain and 0.0367 for Leiden, reflecting its limited ability to



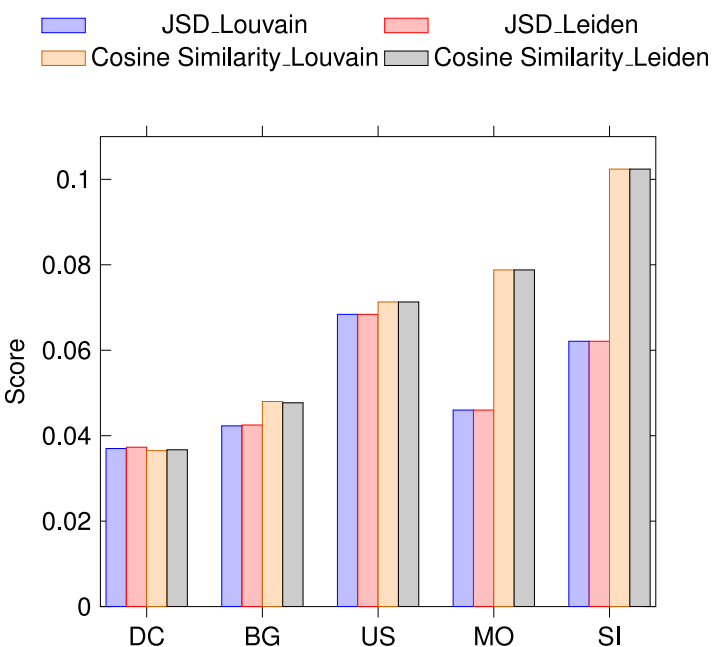
Codes: DC = traditional Direct citation; {BG, US, EX, MO, DI, SI, FW} corresponds to IC_{DC} with citation intents {*Background*, *Uses*, *Extends*, *Motivation*, *Differences*, *Similarities*, *Future work*}, respectively.

Fig. 10 Comparison of Direct citation methods in the NLP domain. **Codes:** DC = traditional Direct citation; {BG, US, EX, MO, DI, SI, FW} corresponds to IC_{DC} with citation intents {*Background*, *Uses*, *Extends*, *Motivation*, *Differences*, *Similarities*, *Future work*}, respectively

capture article relationships. $IC_{DC-Similarities}$ markedly improves textual coherence, achieving coherence using JSD of 0.0621 and the highest score using Cosine Similarity measure of 0.1024 for both clustering methods, obtains enhancement by 0.0659 score compared to baseline model, underscoring its key role in relatedness computation. Besides, $IC_{DC-Motivation}$ demonstrates a solid improvement in textual coherence with a enhancement by 0.0423 score, while $IC_{DC-Uses}$ intent shows gains with JSD-based coherence of 0.0684 and Cosine Similarity-based coherence of 0.0713. Using *Background* intent, $IC_{DC-Background}$ offers a minimal improvement in textual coherence score, remaining the lowest contribution compared to other citation intents.

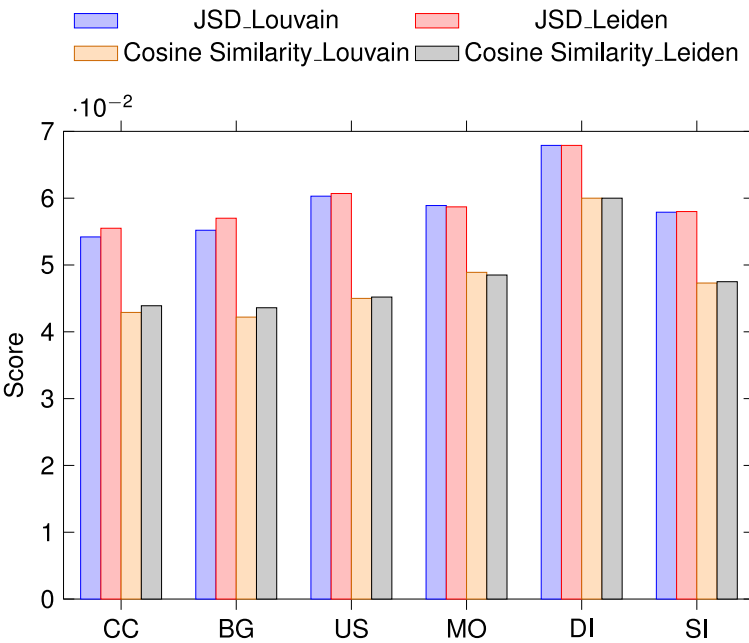
5.2.3 Co-citation

Figure 12 compares Co-citation methods in the NLP domain, excluding *Extends* and *Future work* due to ANRs below the threshold. Overall, the traditional Co-citation method shows the lowest textual coherence performance, with JSD measure of 0.0542 for Louvain and 0.0555 for Leiden, and Cosine Similarity of 0.0429 for Louvain and 0.0439 for Leiden. Moreover, using *Differences* intent, $IC_{CC-Differences}$ significantly outperformed all other methods, achieving the highest textual coherence with JSD (0.0679) and Cosine Similarity (0.0600), enhancing Co-citation network relatedness. Moreover, $IC_{CC-Similarities}$ improves on the baseline, with coherence using JSD of 0.0580 and Cosine Similarity of 0.0475 for Leiden method. Besides, $IC_{CC-Uses}$ intent excels in JSD-based coherence (0.0603 for Louvain, 0.0607 for Leiden), while $IC_{CC-Motivation}$ intent leads the performance in Cosine Similarity-based coherence (0.0489 for Louvain, 0.0485



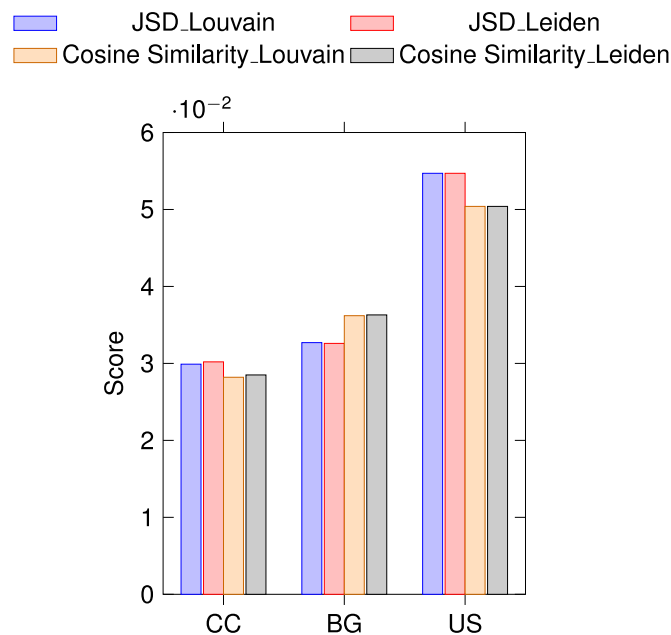
Codes: DC = traditional Direct citation; {BG, US, MO, SI} corresponds to IC_{DC} with citation intents {*Background, Uses, Motivation, Similarities*}, respectively.

Fig. 11 Comparison of Direct citation methods in the Biomedical domain. **Codes:** DC = traditional Direct citation; {BG, US, MO, SI} corresponds to IC_{DC} with citation intents {*Background, Uses, Motivation, Similarities*}, respectively



Codes: CC = traditional Co-citation; {BG, US, MO, DI, SI} corresponds to IC_{CC} with citation intents {*Background, Uses, Motivation, Differences, Similarities*}, respectively.

Fig. 12 Comparison of Co-citation methods in the NLP domain. **Codes:** CC = traditional Co-citation; {BG, US, MO, DI, SI} corresponds to IC_{CC} with citation intents {*Background, Uses, Motivation, Differences, Similarities*}, respectively



Codes: CC = traditional Co-citation; {BG, US} corresponds to IC_{CC} with citation intents {*Background*, *Uses*}, respectively.

Fig. 13 Comparison of Co-citation methods in the Biomedical domain. **Codes:** CC = traditional Co-citation; {BG, US} corresponds to IC_{CC} with citation intents {*Background*, *Uses*}, respectively

for Leiden). Finally, $IC_{CC-Background}$ produces slight improvements in JSD-based coherence scores but performs worse among all Intent-aware methods.

Figure 13 shows the textual coherence results in the Biomedical dataset, including only *Background* and *Uses* intents due to their ANRs exceeding the threshold. Overall, $IC_{CC-Uses}$ consistently outperforms both the traditional Co-citation and $IC_{CC-Background}$ across all metrics, achieving the highest textual coherence with JSD of 0.0547 and Cosine Similarity of 0.0504 for both cluster methods. These scores surpass the coherence score of both our method using *Background* intent and traditional Co-citation.

6 Discussion

6.1 Impact of citation intents on publication relatedness

In this part, we analyze and discuss the impact of citation intents on publication relatedness.

Traditional citation-based methods treat all citations uniformly, overlooking the varied roles of citing papers, such as providing background context, discussing applications, extending methods or tools, offering motivations, or making comparisons. This approach often creates weak or unrelated connections between articles, which bring different rhetorical functions, diminishing the quality of publication relatedness and clustering outcomes. Our experiments demonstrate that these methods result in lower textual coherence compared to Intent-aware approaches, highlighting the need for advanced techniques that consider citation motivations to improve the accuracy and value of scholarly networks.

In contrast, our approach, which categorizes citations by analyzing citation intent, achieves significant improvements in publication relatedness, particularly when leveraging the intents for comparison purposes. The group of Comparison intents, encompassing both *Similarities* and *Differences*, deliver superior performance across most scenarios. This is unsurprising, as both *Similarities* and *Differences* generally signal stronger conceptual proximity between citing and cited papers, as meaningful comparison requires alignment across multiple dimensions of the research process. In such contexts, the citing work explicitly relates alternative approaches that address the same underlying research problem, thereby necessitating shared problem formulations, assumptions, and scope. These comparisons further presuppose consistency in evaluation practices, including comparable metrics and experimental settings, which constrain the surrounding discourse and promote higher topical coherence. From a rhetorical perspective, the citations with comparison purposes position the cited work as an intellectual peer, serving as a reference point against which the citing paper articulates both commonalities and distinctions. Overall, these characteristics explain why comparison-based citation intents consistently yield stronger signals for capturing publication relatedness.

Extends intent fosters robust connection between articles, as the source papers build upon, improves, or modifies concepts, methods, tools from the relevant works. Moreover, intents such as *Motivation*, *Uses*, and *Future work* yield promising results compared to traditional methods. The reason is that *Uses* and *Future work* intents involve leveraging specific tools, methods, or data from a target paper as components within the source paper's context, while *Motivation* intents focus on limited aspects like data, goals, or methods from closely related works. However, these intents are less effective than the group of Comparison intents, because they often impose weaker topical constraints because they typically operate on specific aspects such as: goal, data, ideas, methods, or tools from prior work. These elements usually account for only a limited portion of the overall research context of a scientific paper, resulting in a narrow and localized scope of connection between the citing and cited works. Consequently, the topical overlap induced by these intents tends to be partial rather than global. In contrast, for a comparison to be meaningful, the group of Comparison intents (*Similarities* and *Differences*) tends to require the citing and cited papers to share core aspects such as problem formulation, methodological scope, and evaluative criteria, which enforces stronger topical constraints and sustains a broader shared discourse. Although this finding differs from prior studies, which often emphasize *Uses* and *Extends* as the strongest indicators of conceptual advancement [18, 19] rather than the the group of Comparison intents as in our study, our findings do not negate prior work that emphasizes the importance of intents such as *Uses* or *Extends*. Rather, this study aims to complement existing research by highlighting and systematically analyzing a group of citation intents that have received comparatively less attention in previous studies, such as *Differences* and *Similarities*.

The *Background* intent yields the poorest performance in connecting papers compared to other citation intents due to its inherently superficial and broad nature, which limits the depth of contextual overlap between citing and cited works. Unlike intents such as *Differences*, *Similarities*, *Extends*, *Uses*, *Future Work*, or *Motivation*, which engage with specific aspects of a target paper, such as shared methodologies, tools, data, or research

goals, many *Background* citations serve as introductory or general context, often referencing works from loosely related domains to set the stage for the source paper's narrative. As a result, *Background* intents produce minimal contextual overlap, reducing effectiveness in establishing robust publication relatedness in scholarly networks.

6.2 Robustness and generalizability across multi-label schemes and scientific domains

We first examine the robustness of the proposed approach under the multi-label setting of the MultiCite dataset, where a single citation context may express multiple citation intents simultaneously. This phenomenon does not have any impact on the overall results of the proposed method. Instead, using this dataset allows for a more accurate characterization of the inherent properties of citation contexts, in which a single citation context may convey more than one citation intent. This perspective naturally supports the notion that a paper can cite another work through multiple types of relationships, and that the strength of these relationships may vary depending on the specific intent involved. Such a setting does not negatively affect the model's overall performance; rather, it provides a more comprehensive framework for evaluating inter-paper relatedness by capturing both the diversity of citation relations and their varying degrees of strength.

We further investigate the generalizability of our approach across datasets from different domains, including NLP and Biomedical corpora, which exhibit distinct citation practices and discourse conventions. The specific characteristics of the NLP and Biomedical datasets do not alter the overall results of our approach. While minor quantitative variations across datasets are observed, the main performance trends remain consistent, indicating that the proposed method is not domain-dependent. In particular, across all datasets, approaches that explicitly incorporate citation intent consistently outperform traditional models that do not leverage intent information. Among individual intent categories, *Background* exhibits the lowest effectiveness, reflecting its largely descriptive role in citation contexts. By contrast, intent groups with comparative functions, such as *Differences* and *Similarities*, achieve the strongest performance and consistently surpass intents associated with methodological usage or extension, including *Uses*, *Motivation*, *Extends*, and *Future Work*. These consistent trends across domains further reinforce the robustness and generalizability of the proposed approach. This behavior is reasonable in practice, as citation intents capture rhetorical functions that are intrinsic to scholarly writing, such as background description, comparison (including similarities and differences), methodological use, extension of prior work, and the motivation or justification of research choices. These rhetorical functions are shared across scientific disciplines, which helps explain the consistent performance of our method across domains.

6.3 Transparency, classifier bias, responsible deployment and computation costs

In this section, we consider and discuss issues related to transparency, classifier bias, and responsible deployment, as our work could influence scholarly search engines and knowledge-mapping systems. Besides, we also examine the computational costs of our method in comparison to traditional methods.

6.3.1 Transparency

Our model is designed to be transparent in terms of the induced paper-paper relationships and the intent-level decision process. In the search engines or knowledge-mapping systems, our system provides a clear explanation for how two papers are related, as if they are connected directly (DC), sharing the same cited papers (BC), or sharing the same citing papers (CC), based on the citation intents of citation context. The process of identifying citation intent labels for input citation text is also transparent, with the output of the neural network as the probability of whether the specific intent is chosen. The citation intent that has a probability higher than the threshold will be chosen.

6.3.2 Classifier bias

In this part, we analyze potential classifier bias in our citation intent classification model by examining several major sources of bias. First, data imbalance bias arises from the uneven distribution of training samples across citation intent categories, which may cause the model to favor high-frequency intents. To assess this effect, we provide Table 5 to illustrate the classification performance with the corresponding number of training samples for each citation intent. Overall, while the classifier achieves higher performance for intent categories with more abundant training samples such as *Background*, *Uses*, *Differences*, it nevertheless maintains reasonably strong and stable performance even for low-resource intents such as *Future work*. This observation suggests that the proposed classifier is less sensitive to data imbalance and exhibits limited bias across citation intent categories. Second, we further analyze classifier bias induced by data heterogeneity, including variations in citation styles, and the presence of data noise. Particularly, although the MultiCite dataset contains citation contexts with a wide variety of citation styles, the model continues to achieve strong and stable performance, suggesting limited bias with respect to citation style variation.

Additionally, regarding data noise issue, the use of SciBERT, a pretrained language model as the word encoding layer to extract contextual representations helps mitigate the impact of data noise and reduces noise-induced bias. Overall, these results suggest that the classifier is not overly influenced by data imbalance or heterogeneous data, making it a reliable component for real-world citation intent analysis.

6.3.3 Responsible deployment

From a responsible deployment perspective, it is important to clearly delineate the scope of the proposed similarity measure. Our approach focuses exclusively on citation-based similarity, capturing relationships derived from shared citation intents rather than text-based semantic similarity. Accordingly, the resulting similarity scores should be interpreted within this methodological scope, rather than as a similarity measure that simultaneously captures both citation-based and text-based similarity. Additionally, similarity scores derived from different citation intent categories should be interpreted separately to preserve a clear understanding of the origin of the citation-based relationship and the relative strength associated with each intent-specific connection. From this perspective, intent-based similarity should be treated as complementary information and integrated with other bibliographic, textual, or domain-specific signals, rather than used in isolation when interpreting paper-paper relatedness.

6.3.4 Computational costs

In this part, we discuss about the comparison of computational costs between IC and traditional methods. Specifically, after citation intent labels are obtained for all citation contexts, the construction of intent-aware BC and CC follows the same counting and aggregation procedures as their traditional counterparts, with the only difference being the inclusion of intent-based filtering. Consequently, the computational complexity of Bibliographic coupling and Co-citation computation remains essentially unchanged. Any additional computational overhead therefore arises solely from the citation intent classification step, which is performed once as an offline preprocessing stage using a SciBERT-based model. Although the use of a SciBERT-based classifier introduces an additional preprocessing cost, the resulting citation intent annotations can be reused across multiple downstream analyses, so that this cost is incurred only once and does not accumulate with repeated use. Overall, the additional computational cost introduced by the proposed framework remains limited and acceptable.

6.4 Hard and soft citation intent matching

In this part, we discuss two alternative strategies for incorporating citation intent information into publication relatedness modeling: hard intent matching and soft (weighted) intent matching. Our proposed approach adopts a hard citation intent matching strategy, in which citation links are retained only when the citing and cited publications share the same intent. While this design choice effectively filters out superficial or semantically weak connections by enforcing intent consistency (e.g., a paper cited for *Background* versus one cited for *Uses*), it may overlook certain latent relationships between papers associated with closely related but distinct citation intents (e.g., *Motivation* and *Extends*).

An alternative solution is to consider a soft or weighted formulation for citation intent matching. Under such a formulation, citation intents are no longer treated as strictly equivalent or unrelated, but are instead assigned graded similarity values reflecting partial functional overlap, for example by defining a similarity matrix between citation intent categories. Rather than enforcing exact intent agreement, citation links associated with different intents may contribute to publication relatedness with varying strengths, depending on the degree of rhetorical similarity between the corresponding intent categories. Such a soft weighting mechanism may be beneficial in cases where intent boundaries are conceptually close or exhibit gradual transitions between rhetorical functions.

Although a soft intent matching strategy is theoretically appealing and promising, implementing such an approach remains challenging. Particularly, to date, prior studies have not provided a precise or widely accepted assessment of the degree of relatedness between different citation intent types. As a consequence, establishing a theoretical framework for intent similarity would require introducing additional assumptions regarding how rhetorical functions relate to one another. Even when relative relationships are available, for example, knowing that intent X is closer to intent Y than to intent Z, there is still no principled mechanism for translating such ordinal relationships into a real-valued similarity matrix suitable for modeling purposes.

Compared to soft intent matching, hard citation intent matching offers a more straightforward and easily applicable solution. It does not require the construction of a new theoretical framework to quantify the relatedness between citation intent types, nor

does it rely on additional mechanisms to map relative or qualitative intent similarities into real-valued relevance scores. As a result, the hard formulation avoids introducing extra modeling complexity and can be readily applied without making further assumptions about inter-intent relationships. For these reasons, hard citation intent matching is adopted as the preferred choice in this study.

6.5 Broader implications

Beyond paper-level relationship modeling, the proposed approach has broader implications for bibliometric analysis. By incorporating citation intent as a function-aware constraint, our method moves beyond traditional count-based or co-occurrence-driven analyses and enables a more semantically grounded understanding of scholarly relationships.

More broadly, the proposed approach can serve as a methodological foundation for a class of context-aware methods that exploit functional properties derived from citation contexts for modeling relationships between scientific papers, such as citation polarity. By incorporating contextual characteristics of citations, this framework enables deeper investigation into the nature of inter-paper relationships based on the citation context-based analysis.

Beyond the scope of publication relatedness task, our approach naturally suggests the application of citation intent to other bibliometric tasks, such as assessing the influence or importance of scientific publications, constructing scientific knowledge maps, and analyzing the evolutionary trajectories of scholarly knowledge.

Moreover, this approach not only opens new research opportunities for applying citation intent in paper-paper relationship analysis, but also provides a foundation for extending intent-aware modeling to other scholarly entities at multiple levels such as authors, journals, and conferences. By aggregating intent-specific citation patterns at different levels of analysis, future work may characterize the functional roles of authors, venues, and research communities. As a result, it motivates new directions for modeling and analyzing complex relationships among scholarly entities based on information derived from citation intents, thereby providing a more comprehensive and holistic understanding of the nature of their interrelationships. Such extensions would enable more nuanced and comprehensive bibliometric analyses, moving beyond purely structural indicators toward representations that reflect the diverse roles underlying scholarly interactions.

7 Conclusion and future works

7.1 Conclusion

In this paper, we introduce an innovative Intent-aware citation-based method (IC) that enhances the measurement of publication relatedness by incorporating citation intent. Our approach begins by defining seven distinct citation intent relation types based on the MultiCite dataset. We then develop a Citation intent classifier for scientific publications, to accurately predict these citation intent categories. By leveraging these predicted intents, IC compute relatedness scores between articles for each citation-based relation and construct Intent-aware citation networks. We evaluate our proposed method on two diverse datasets, such as NLP publications and Biomedical publications. By employing clustering algorithms on citation networks and measuring textual coherence with

Jensen-Shannon divergence and Cosine Similarity, we showed that our method, IC significantly outperforms traditional citation-based approaches. Our findings reveal that using the citation intents with comparison purposes, encompassing both *Differences* and *Similarities*, consistently yields superior performance across all citation scenarios. In contrast, the group of *Extends*, *Motivation*, *Uses*, and *Future work* intents demonstrates comparatively lower effectiveness in building scholarly connections, though they still contribute meaningfully to the citation network. Notably, using *Background* intent yields the weakest links within the Intent-aware framework, underscoring its limited capacity to drive strong publication relatedness. Beyond paper-paper relationship modeling, our approach has broader impact for bibliometric analysis by providing a principled, intent-aware perspective on how citation contexts functionally shape scholarly connections.

7.2 Future works

We explicitly describe several promising avenues for future work, including the integration of content-based signals from citation contexts and the exploration of hierarchical citation intent models to capture multi-level citation intent relationships. Particularly, regarding leveraging the content-based signals, future work may replace the current hard within-intent matching with a soft, content-aware similarity by computing cosine similarity between embedding representations of citation contexts after intent filtering. This formulation would enable the model to capture the semantic relatedness even among citations sharing the same functional roles. Regarding using hierarchical citation intent, future research may investigate whether a hierarchical structure of citation intents exists, in which latent relationships among different citation intents can be identified. Under this hierarchical formulation, paper-paper similarity could be computed based on intent relationships defined within the same level of abstraction or across different levels of the hierarchy, enabling the framework to capture both coarse-grained and fine-grained functional relatedness.

Author contributions

Tuan Anh Phan: Conceptualization, Methodology, Implementation, Formal analysis, Writing and Editing. Jason J. Jung: Conceptualization, Methodology, Supervision, Formal analysis, Writing and Editing.

Funding

No funding was received for this work.

Data availability

The MultiCite dataset analysed during the current study is available in the MultiCite repository, <https://github.com/allenai/multicite>. The NLP dataset analysed during the current study is available in the ACL-anthology-corpus repository, <https://github.com/shauryr/ACL-anthology-corpus>. The Biomedical dataset analysed during the current study is available in a public repository, titled "Clustering Scientific Document Based on an Extended Citation Model", <https://doi.org/10.6084/m9.figshare.7634489>.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 25 September 2025 / Accepted: 9 February 2026

Published online: 16 February 2026

References

1. Ortuño FM, Rojas I, Andrade-Navarro MA, Fontaine J-F. Using cited references to improve the retrieval of related biomedical documents. *BMC Bioinform.* 2013;14(1):113.
2. Janssens ACJ, Gwinn M, Brockman JE, Powell K, Goodman M. Novel citation-based search method for scientific literature: a validation study. *BMC Med Res Methodol.* 2020;20(1):25.
3. Cole V, Boutet M. Researchrabbit. The Journal of the Canadian Health Libraries Association. 2023;44(2):43.
4. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. Science mapping software tools: review, analysis, and cooperative study among tools. *J Am Soc Inform Sci Technol.* 2011;62(7):1382–402.
5. Chen C. Science mapping: a systematic review of the literature. *J Data Inform Sci.* 2017;2(2):1.
6. Waltman L, Van Eck NJ. A new methodology for constructing a publication-level classification system of science. *J Am Soc Inform Sci Technol.* 2012;63(12):2378–92.
7. Waltman L, Boyack KW, Colavizza G, Eck NJ. A principled methodology for comparing relatedness measures for clustering publications. *Quant Sci Stud.* 2020;1(2):691–713.
8. Martyn J. Bibliographic coupling. *J Doc.* 1964;20(4):236–236.
9. Small H. Co-citation in the scientific literature: a new measure of the relationship between two documents. *J Am Soc Inform Sci.* 1973;24(4):265–9.
10. Fujita K, Kajikawa Y, Mori J, Sakata I. Detecting research fronts using different types of weighted citation networks. *J Eng Tech Manage.* 2014;32:129–46.
11. Hamedani MR, Kim S-W, Kim D-J. Simcc: a novel method to consider both content and citations for computing similarity of scientific papers. *Inf Sci.* 2016;334:273–92.
12. Zhang S, Xu Y, Zhang W. Clustering scientific document based on an extended citation model. *IEEE Access.* 2019;7:57037–46.
13. Sjögård P, Ahlgren P. Normalization of direct citations for clustering in publication-level networks: evaluation of six approaches. *Scientometrics.* 2024;129(3):1949–68.
14. Goltaji M, Abbaspour J, Jowkar A, Fakhrahmad SM. Comparison of text-based and linked-based metrics in terms of estimating the similarity of articles. *J Librariansh Inf Sci.* 2024;56(3):760–72.
15. Bascur JP, Verberne S, Eck NJ, Waltman L. Which topics are best represented by science maps? an analysis of clustering effectiveness for citation and text similarity networks. *Scientometrics.* 2025;130(2):1181–99.
16. Lauscher A, Ko B, Kuehl B, Johnson S, Cohan A, Jurgens D, Lo K. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In: Carpuat M, Marneffe M-C, Meza Ruiz IV, editors. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. Seattle. 2022. pp. 1875–1889. <https://doi.org/10.18653/v1/2022.naacl-main.137>. <https://aclanthology.org/2022.naacl-main.137/>.
17. Lahiri A, Sanyal DK, Mukherjee I. Citeprompt: using prompts to identify citation intent in scientific papers. In: 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL). 2023. pp. 51–55. IEEE.
18. Valenzuela M, Ha V, Etzioni O. Identifying meaningful citations. In: *AAAI Workshop: scholarly Big Data*, vol. 15, 2015. pp. 13.
19. Hassan S-U, Akram A, Haddawy P. Identifying important citations using contextual information from full text. In: 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL). 2017. pp. 1–8. IEEE.
20. Qayyum F, Afzal MT. Identification of important citations by exploiting research articles' metadata and cue-terms from content. *Scientometrics.* 2019;118(1):21–43.
21. An X, Sun X, Xu S, Hao L, Li J. Important citations identification by exploiting generative model into discriminative model. *J Inf Sci.* 2023;49(1):107–21.
22. James M, Palakkat V, Jones GJ. Identifying influential citations in scientific papers. In: 2023 31st Irish Conference on Artificial Intelligence and Cognitive Science (AICS), 2023. pp. 1–4. IEEE.
23. Boyack KW, Klavans R. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *J Assoc Inf Sci Technol.* 2010;61(12):2389–404. <https://doi.org/10.1002/ASI.21419>.
24. Sjögård P, Ahlgren P. Seed-based information retrieval in networks of research publications: evaluation of direct citations, bibliographic coupling, co-citations, and pubmed-related article score. *J Am Soc Inf Sci.* 2024;75(13):1453–65.
25. Wang X, Sun Y. Research front and future prospect: a bibliometric study of metaphor comprehension. *SAGE Open.* 2024;14(2):21582440241241430.
26. Xie Q, Waltman L. A comparison of citation-based clustering and topic modeling for science mapping. *Scientometrics.* 2025;1–26.
27. Ahlgren P, Chen Y, Colliander C, Eck NJ. Enhancing direct citations: a comparison of relatedness measures for community detection in a large set of pubmed publications. *Quant Sci Stud.* 2020;1(2):714–29.
28. Hassan S-U, Imran M, Iqbal S, Aljohani NR, Nawaz R. Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics.* 2018;117(3):1645–62.
29. Teufel S, Siddharthan A, Tidhar D. Automatic classification of citation function. In: *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. 2006. pp. 103–110.
30. Abu-Jbara A, Ezra J, Radev D. Purpose and polarity of citation: towards nlp-based bibliometrics. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2013. pp. 596–606.
31. Aljohani NR, Fayoumi A, Hassan S-U. An in-text citation classification predictive model for a scholarly search system. *Scientometrics.* 2021;126(7):5509–29.
32. Hannachi F, Hammache A, Saidani F. Identifying important citation through cascading models and nlp techniques. In: 2024 International Conference of the African Federation of Operational Research Societies (AFROS). 2024. pp. 1–6. IEEE.
33. Iqbal A, Shahid A, Roman M, Afzal MT, Hassan UU. Optimising window size of semantic of classification model for identification of in-text citations based on context and intent. *PLoS ONE.* 2025;20(3):0309862.
34. Houghton AM, Quintero PA, Perkins DL, Kobayashi DK, Kelley DG, Marconcini LA, et al. Elastin fragments drive disease progression in a murine model of emphysema. *J Clin Invest.* 2006;116(3):753–9.
35. Assi HI, Kamphorst AO, Moukalled NM, Ramalingam SS. Immune checkpoint inhibitors in advanced non-small cell lung cancer. *Cancer.* 2018;124(2):248–61.

36. Jackson EL, Willis N, Mercer K, Bronson RT, Crowley D, Montoya R, et al. Analysis of lung tumor initiation and progression using conditional expression of oncogenic k-ras. *Genes Dev.* 2001;15(24):3243–8.
37. Tran E, Ahmadvadeh M, Lu Y-C, Gros A, Turcotte S, Robbins PF, et al. Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science.* 2015;350(6266):1387–90.
38. Sade-Feldman M, Yizhak K, Bjorgaard SL, Ray JP, De Boer CG, Jenkins RW, et al. Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell.* 2018;175(4):998–1013.
39. Angermueller C, Clark SJ, Lee HJ, Macaulay IC, Teng MJ, Hu TX, et al. Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity. *Nat Methods.* 2016;13(3):229–32.
40. Pelka K, Hofree M, Chen JH, Sarkizova S, Pirl JD, Jorgji V, et al. Spatially organized multicellular immune hubs in human colorectal cancer. *Cell.* 2021;184(18):4734–52.
41. Lieberman J. Granzyme a activates another way to die. *Immunol Rev.* 2010;235(1):93–104.
42. Peng Y-M, Garde MD, Cheng K-F, Baars PA, Remmerswaal EB, Lier RA, et al. Specific expression of gpr56 by human cytotoxic lymphocytes. *J Leukoc Biol.* 2011;90(4):735–40.
43. Baixauli F, Martín-Cófreces NB, Morlino G, Carrasco YR, Calabia-Linares C, Veiga E, et al. The mitochondrial fission factor dynamin-related protein 1 modulates t-cell receptor signalling at the immune synapse. *EMBO J.* 2011;30(7):1238–50.
44. Wang H, Guo Y, Dong Z, Li T, Xie X, Wan D, et al. Differential u2af1 mutation sites, burden and co-mutation genes can predict prognosis in patients with myelodysplastic syndrome. *Sci Rep.* 2020;10(1):18622.
45. Zheng C, Fass JN, Shih Y-P, Gunderson AJ, Silva NS, Huang H, et al. Transcriptomic profiles of neoantigen-reactive t cells in human gastrointestinal cancers. *Cancer Cell.* 2022;40(4):410–23.
46. Hanada K-I, Zhao C, Gil-Hoyos R, Gartner JJ, Chow-Parmer C, Lowery FJ, et al. A phenotypic signature that identifies neoantigen-reactive t cells in fresh human lung cancers. *Cancer Cell.* 2022;40(5):479–93.
47. Marwitz S. Lung cancer subtypes generate unique immune responses. *Breathe* 2017;13(2).
48. Beltagy I, Lo K, Cohan A. Scibert: a pretrained language model for scientific text. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019. Association for Computational Linguistics. 2019. pp. 3613–3618. <https://doi.org/10.18653/v1/D19-1371>.
49. Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y, editors. 3rd International Conference on Learning Representations, ICLR 2015. San Diego. May 7–9. 2015. Conference Track Proceedings 2015. [arxiv:1412.6980](https://arxiv.org/abs/1412.6980).
50. Rohatgi S. ACL Anthology Corpus with Full Text. Github 2022. <https://github.com/shauryr/acl-anthology-corpus>.
51. Lin J. Divergence measures based on the shannon entropy. *IEEE Trans Inf Theory.* 1991;37(1):145–51. <https://doi.org/10.1109/9.1861115>.
52. Leicht EA, Newman ME. Community structure in directed networks. *Phys Rev Lett.* 2008;100(11):118703.
53. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp.* 2008;2008(10):10008.
54. Traag VA, Waltman L, Van Eck NJ. From louvain to leiden: guaranteeing well-connected communities. *Sci Rep.* 2019;9(1):1–12.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.