



# Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications

Jianguen He\* and Chaomei Chen\*

Department of Information Science, College of Computing and Informatics, Drexel University, Philadelphia, PA, United States

## OPEN ACCESS

### Edited by:

Marc Bertin,  
Claude Bernard University Lyon 1,  
France

### Reviewed by:

Nicolás Robinson-García,  
Universitat Politècnica de València,  
Spain

Benjamin Vargas-Quesada,  
Universidad de Granada, Spain

Jiang Wu,  
Wuhan University, China

### \*Correspondence:

Jianguen He  
jianguen.he@drexel.edu  
Chaomei Chen  
chaomei.chen@drexel.edu

**Received:** 18 February 2018

**Accepted:** 27 August 2018

**Published:** 19 September 2018

### Citation:

He J and Chen C (2018) Temporal  
Representations of Citations for  
Understanding the Changing Roles of  
Scientific Publications.  
*Front. Res. Metr. Anal.* 3:27.  
doi: 10.3389/frrma.2018.00027

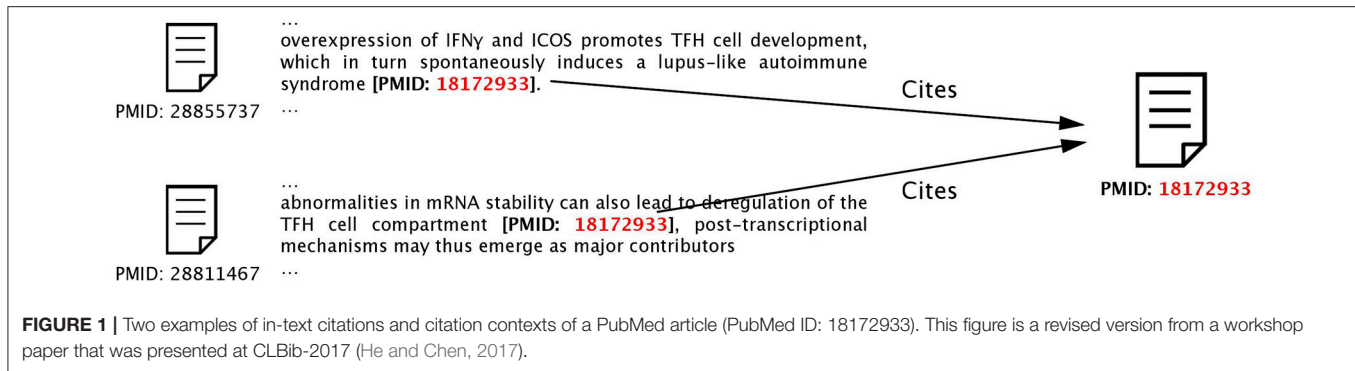
Researchers may describe different aspects of past scientific publications in their publications and the descriptions may keep changing in the evolution of science. The diverse and changing descriptions (i.e., citation contexts) on a publication characterize the impact and contributions of the past publication. In this article, we aim to provide an approach to understanding the changing roles of a publication characterized by its citation contexts in the full text of publications. We proposed approaches for representing the changing citation contexts of cited publications in different periods as sequences of vectors by training temporal embedding models. We can utilize the temporal representations to quantify how much the roles of publications changed and interpret how they changed. We also evaluated the performance of three ways of constructing citation contexts for representation learning. Our study in the biomedical domain shows that our metric on the changes of publication roles is stable at the group level but it can account for the variation of individual publications.

**Keywords:** in-text citation, citation context, embedding learning, citation analysis, document representation, full-text literature

## 1. INTRODUCTION

Although the content of a scientific publication cannot be changed once it was published, how other researchers cite and evaluate the publication may keep changing. The actual scientific contributions and impact of a specific publication are changing within the evolving intellectual spaces constructed by other publications. Besides the role of a publications are ever-changing, the role of may be complex because of the varied contributions made the publication, especially the publications contributed to interdisciplinary or fundamental research topics. The changing and complex roles of cited publications can be characterized by their citations and citation contexts.

Both citation network and citation context (i.e., the sentences containing in-text citations) can be utilized for analyzing scientific publications (Elkiss et al., 2008). A relevant intellectual structure characterized by citation network is commonly used as a foundation for analyzing the role of a publication played in scientific dynamics, such as identifying the place where the analyzed publication is in the intellectual structure (Orosz et al., 2016) or the structural alteration caused by the publication (Chen, 2012). The text of citation contexts were also used to characterize publications for various applications, such as publication summarization (Qazvinian et al., 2010), survey article generation (Mohammad et al., 2009), and information retrieval (Huang et al., 2015). Quantitative metric for quantifying the role changes of publications can be derived from



citation network analysis, but interpreting the changes is not straightforward which always relied on techniques of text mining and visual analytics. While analyzing the text of citation contexts naturally has interpretable results but a unified quantitative measurement is challenging to be built on the unstructured textual data.

In this article, we proposed methods for learning temporal representations of in-text citations of publications by word embedding models, which can be used to characterize and analyze the changing roles of the publications. The in-text citations of publications are the citations referred to this publication within the full text of publications cited this publication; The text around the in-text citation is the citation context text (see **Figure 1** for examples). We proposed and compared different ways of constructing the citation contexts for representation learning. Based on the temporal representations of citations, we introduced a simple method to quantify the role changes of publications characterized by their citation contexts. We also analyzed the distribution of change scores and described applications of how to identify and interpret the changes by making use of the embedding representations.

## 2. RELATED WORK

Due to the availability full-text data of scientific articles, such as PubMed Central, many citation-based studies went beyond the article metadata and citation links. The proximity of citations was combined with co-citation analysis to provide co-citation networks at multiple levels of granularity (Liu and Chen, 2012) or to identify related work (Gipp and Beel, 2009). Citation contexts also has been utilized to improve co-citation network analysis (Callahan et al., 2010; Small and Klavans, 2011; Boyack et al., 2013) and enhance the application of direct citation network (Sugiyama and Kan, 2013). Some studies emphasized the literal features of citation by analyzing the citation context, such author's reason for citing (Teufel et al., 2006) and sentiment of citation (Small, 2011). Besides, various applications based on citation context have been developed, such as information retrieval (Eto, 2013; Liu S. et al., 2014a) and article recommendation (He et al., 2010; Liu X. et al., 2014b).

More recently, embedding learning techniques were employed in representing key elements of scientific knowledge, such as publications (Ganguly and Pudi, 2017),

authors (Ganesh et al., 2016), citations (Berger et al., 2017), and research topics (He and Chen, 2018). Paper2vec (Ganguly and Pudi, 2017) leveraged both citation networks and textual information of publications to represent a publication, but the textual information they used is the full text of publications which is the description from authors of publication rather than scientific communities. Another study also named *Paper2vec* (Tian and Zhuo, 2017) focused on utilizing neighbor nodes of publications in citation network to represent the publications. *Cite2vec* (Berger et al., 2017) represented publications by using their citation contexts as ours and provided visualization for exploration, while the temporal feature of citation contexts is ignored. In our study, we emphasize representing the changes of publications characterized by the citation contexts of the publications over time.

## 3. METHODS

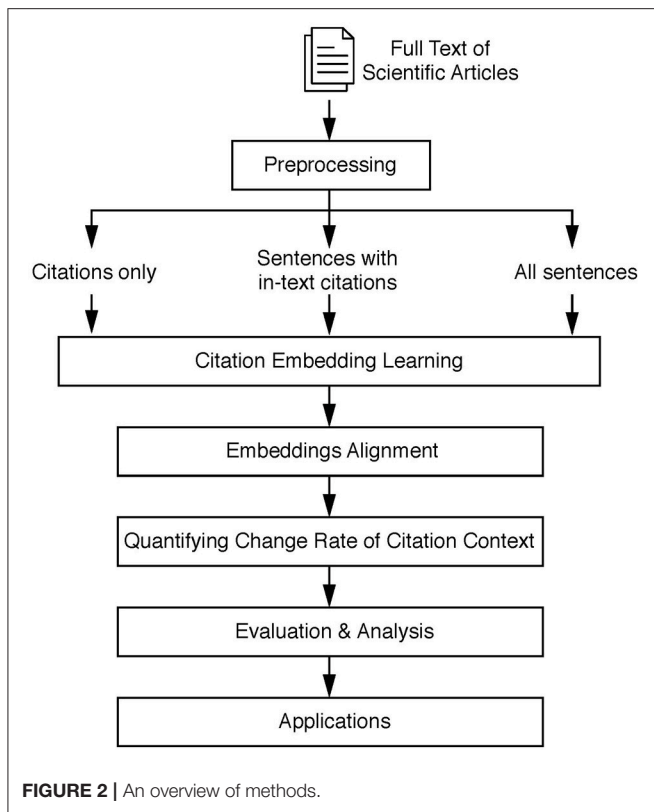
In this section, we describe how we train temporal citation embedding models, which includes data preprocessing, constructing citation contexts, embedding model training over periods, and embedding model alignment. We also present our approach to quantifying the role changes of publications. **Figure 2** describes the overview of our methods.

### 3.1. Data and Preprocessing

The dataset we use for training is the PubMed Central Open Access Subset (PMC OAS), which is an open access XML formatted full-text document repository from biomedicine and life sciences maintained by the U.S. National Library of Medicine (NLM)<sup>1</sup>. We can parse in-text citations and citation contexts from PMC OAS because of the well-structured XML files. In this study, we trained embedding models by documents published from 2007 to 2016 in PMC OAS, which comprises 1,361,455 full-text scientific publications.

To train the citation embeddings, we need to use a unique identifier to indicate a publication in full text. Many references cited by publications in the PMC OAS have unique publication identifiers such as PubMed ID (PMID), PubMed Central ID (PMCID), and Digital Object Identifier (DOI). However, many cited references don't have unique identifiers. We assign unique

<sup>1</sup><http://www.ncbi.nlm.nih.gov/pmc/>



identifiers for these references by using their metadata in the form of 'FA\_VE\_YR\_VO\_FP' where FA is the first author's first name and last name, VE is the name of venue (journal or conference proceeding), YR is the year of the publication date, VO is the volume number, and FP is the first page number of the publication. If a cited reference has neither a standard identifier nor identifiable metadata for constructing a unique identifier, the reference will be ignored in this study.

About 5.5% references cannot be identified in the PMC OAS dataset. Excluding these unidentifiable references and their in-text citations may have effects on learning representations of citations, because the in-text citations are a part of citation contexts for learning. However, the effects may not be significant. First, both words and in-text citations are the citation contexts for learning, but words constitute the major part of the citation contexts. Second, other identified in-text citations and words can work as substitutes to provide effective contextual information to diminish the effects.

To facilitate citation embedding learning, we convert sentences with in-text citation XML tags into plain text. We only retain text and in-text citations by removing XML tags or converting XML tags into text. We replace the in-text citation tag `<xref ref-type="bibr"></xref>` by using a unique identifier. It is worth noting the various usages of `<xref>` tag in the XML full text. For example, a single `<xref>` may refer to a group of citations (see **Figure 3A**) and some in-text citations are not explicit in XML files (the purple citation identifier in **Figure 3B** is the citation omitted in the XML file).

Since our embedding learning method learns the representation of a citation by capturing the context of the citation within its sentence, we need sentence tokenizer to segment the full text into sentences. Then, we conduct a series of preprocessing by using *NLPre*<sup>2</sup>, including dash removal, capitalization normalization, and replacing phrase from Medical Subject Headings (MeSH) dictionary.

### 3.2. Constructing Citation Contexts

We use three methods to construct citation contexts for embedding learning. They retain different context information for learning citation representations. To illustrate the methods, we denote a paragraph  $p$  in a scientific publication as two sets of sentences  $S = \{s_1, \dots, s_i, \dots, s_m\}$  and  $SC = \{sc_1, \dots, sc_j, \dots, sc_n\}$ , where  $s_i$  is a sentence without any in-text citation and  $sc_j$  is a sentence with a set of in-text citations  $C_j = \{c_1^j, c_2^j, \dots, c_{k_j}^j | k_j \geq 1\}$ .

- **CITATION\_ONLY.** We only use citation identifiers for embedding learning. Since a single sentence usually has very few in-text citations, we use a sequence of citation identifiers in a paragraph as an input record. One input sequence that can be derived from paragraph  $p$  for **CITATION\_ONLY** is  $\{c_1^1, \dots, c_{k_1}^1, \dots, c_1^n, \dots, c_{k_n}^n\}$ .
- **WITH\_CITATION.** We use sentences as input, but only sentences with at least one in-text citation are used.  $m$  input sequences that can be derived from paragraph  $p$  are  $sc_1, sc_2, \dots, sc_n$ .
- **FULL\_SET.** All sentences in the full text of articles are used for embedding learning.  $m+n$  input sequences that can be derived from paragraph  $p$  are  $s_1, \dots, s_m, sc_1, \dots, sc_n$ .

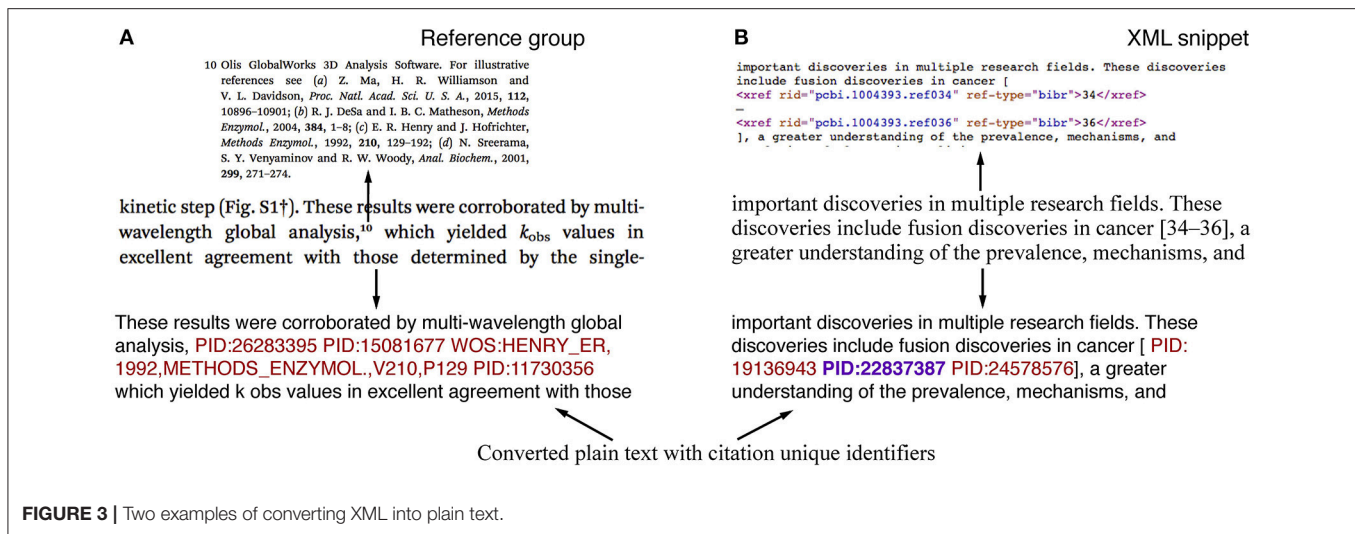
### 3.3. Embedding Learning Methods

We build citation embeddings for understanding how researchers described cited publications. Word embedding techniques were proved to be able to capture semantic and syntactic effectively (Mikolov et al., 2013). We use skip-gram with negative sampling (SGNS) to learn citation embedding based on the context words of citation in the full text of publications. Given a citation or a phrase  $w_i$  in training dataset, skip-gram maps it into a continuous representation vector  $\mathbf{w}_i$ .  $\mathbf{w}_i$  is used to predict the context words of  $w_i$ . The objective of skip-gram is to maximize the log probability:

$$\frac{1}{T} \sum_{i=1}^T \sum_{j=-c}^{c} \log p(w_j | w_i) \quad (1)$$

where  $T$  is the occurrence of each word or citation in the training data,  $c$  is the window size of context and  $w_j$  is the context of  $w_i$ . Negative sampling builds "negative" context words for each  $w_i$  to accelerate the training procedure. We separately constructed citation embeddings from publication text data for each period by SGNS algorithm. We used the implementation of word2vec provided by gensim (Řehůřek and Sojka, 2010) for embedding learning. We empirically set embedding length as 100, negative sampling size as 5, and the number of iteration

<sup>2</sup><https://pypi.python.org/pypi/nlpre>



as 5. However, we set different window size for each type of citation context. For **ONLY\_CITATION**, we set a relatively small window size 3 because of a small length of input sequences and technical practices of training word embeddings of short text. For **WITH\_CITATION** and **FULL\_SET**, we set a larger window size 10. Many in-text citations were placed at the end of sentences, so learning model with a large window size can capture essential context information.

### 3.4. Temporal Embedding Alignment

Since our embedding models are constructed separately for different periods, the models are in different vector space because of differences in stochastic initialization of the weights of the neural network in SGNS algorithm. We need to align the models for different periods into the same coordinate axes to compare citation representations overtime and quantify the citation changes of articles. Following the method proposed by Hamilton et al. (2016), we use orthogonal Procrustes to align the learned embeddings. Defining  $\mathbf{W}^{(t)} \in \mathbb{R}^{d \times |v|}$  as the matrix of word embeddings learn at period  $t$ , we align across time periods while preserving cosine similarities by optimizing

$$\mathbf{R}^{(t)} = \arg \min_{\mathbf{Q}^T \mathbf{Q} = \mathbf{I}} \left\| \mathbf{Q} \mathbf{W}^{(t)} - \mathbf{W}^{(t+1)} \right\|_F \quad (2)$$

with  $\mathbf{R}^{(t)} \in \mathbb{R}^{d \times d}$ . The alignment is performed in an iterative fashion, i.e.,  $(\mathbf{W}^{(1)}, \mathbf{W}^{(2)}), (\mathbf{W}^{(2)}, \mathbf{W}^{(3)}), \dots, (\mathbf{W}^{(T-1)}, \mathbf{W}^{(T)})$  where  $\mathbf{W}^{(t)}$  is the aligned matrix of word embeddings at  $t$ , an alignment of  $(\mathbf{W}^{(t-1)}, \mathbf{W}^{(t)})$  produces an aligned matrix  $\mathbf{W}^{(t)}$ , and  $T$  is the last time-period.

### 3.5. Quantify the Changes of Citation Contexts

The representations of citations can be compared over periods after aligning the citation embeddings over time. The difference between representations of a cited article over periods can be utilized to quantify the change rate of the article's citation

contexts. We measure the difference based on commonly used cosine similarity. Therefore, we quantify the citation change rate of a cited article  $ca$  occurred at  $t$  as

$$\text{Change}^t(ca) = 1 - \cos\_sim(\mathbf{w}_{ca}^t, \mathbf{w}_{ca}^{t-1}) \quad (3)$$

where  $\mathbf{w}_{ca}^t$  is the vector representation of article  $ca$  at  $t$  derived from the citation contexts of  $ca$  at  $t$ .

### 3.6. Evaluation Metric Based on MeSH

We can train the citation embeddings by using three types of citation context described in section 3.2. To evaluate their ability to quantify the change rates of citations of cited articles, we propose an evaluation metric based on MeSH and evaluated the three types of citation context.

We use MeSH to derive an implicit gold standard concerning the topical changes of a publication's citations. Most of publications in this MEDLINE/PubMed (88.25%<sup>3</sup>) were manually assigned a set of descriptors from MeSH by biomedical experts at the U.S. National Library of Medicine (NLM). Similar to multiple prior studies on measuring document similarity (Zhu et al., 2009; Gipp et al., 2015), We view MeSH indexing as an accurate topical description of biomedical articles. The assigned MeSH descriptors of a article collection can describe the topical information of the article collection, so the change over time of assigned MeSH descriptors of the collection of articles cited a certain article can reflect the topical change of the article's citations. Although our temporal citation representation is not designed for representing topical change, a good representation should reflect the topic change as well. Thus, we build an evaluation metric based on the MeSH indexing.

We create the evaluation metric by following the approaches used by CITREC (Gipp et al., 2015) which is evaluation framework for citation-based and text-based similarity measures of documents. However, the evaluation metric proposed by CITREC is for document similarity rather than citation change

<sup>3</sup>[https://www.nlm.nih.gov/bsd/licensee/2017\\_stats/2017\\_LO.html](https://www.nlm.nih.gov/bsd/licensee/2017_stats/2017_LO.html)



of a cited article. Thus, we modified the approaches of CITREC and proposed a metric for measuring the citation change as our evaluation metric. At first, we measure the similarity of MeSH descriptors by utilizing the tree-like structure of MeSH thesaurus. Then, we measure the topical change over time of citations based on the topical information derived from the assigned MeSH descriptors of the citations.

A MeSH descriptor may have multiple tree numbers, which means a descriptor can occur multiple times within the tree structure of MeSH thesaurus. We view the tree numbers as different concepts. To measure the similarity of descriptors, we need to measure the similarity of the concepts behind the descriptors at first. The basic idea of measuring the similarity of two concepts  $c$  and  $c'$  is that the similarity reflects the information they have in common (Gipp et al., 2015). We use the assessment of information content (IC) proposed by Resnik (1995) to quantify the common information of concepts. We quantify information content  $IC$  of a concept  $c$  by a negative log-likelihood function as

$$IC(c) = -\log \frac{1 + s(c)}{N} \quad (4)$$

where  $s(c)$  is the number of concepts subsumed to concept  $c$  and  $N$  is the total number of concepts in the MeSH thesaurus ( $N = 58,760$ ). The common information content of two concepts  $c$  and  $c'$  can be represented as information content of their closest subsuming concept  $c_s(c, c')$ . Then, we calculate the similarity of  $c$  and  $c'$  using Lin's generic similarity measure (Lin et al., 1998) as

$$\text{sim}(c, c') = \frac{2 \times IC(c_s(c, c'))}{IC(c) + IC(c')} \quad (5)$$

To measure the similarity of two MeSH descriptors  $m$  and  $m'$ , we compare the sets of the descriptors' concepts  $C$  and  $C'$ . We use the average maximum match, a similarity measure proposed by Zhu et al. (2009), to calculate the similarity of two MeSH descriptors  $m$  and  $m'$  as

$$\text{sim}(m, m') = \frac{\sum_{c \in C} \max_{c' \in C'} \text{sim}(c, c') + \sum_{c' \in C'} \max_{c \in C} \text{sim}(c, c')}{|C| + |C'|} \quad (6)$$

The similarity of citations of two cited articles  $ca$  and  $ca'$  is determined by the similarity of two sets of articles  $D = \{d | d \text{ cited } ca\}$  and  $D' = \{d' | d' \text{ cited } ca'\}$ . We use the average maximum match between the two sets of MeSH descriptors  $M$  and  $M'$  assigned to citing articles in  $D$  and  $D'$  respectively to measure the citation similarity of  $ca$  and  $ca'$  as

$$\begin{aligned} \text{citation-sim}(ca, ca') &= \text{sim}(D, D') = \text{sim}(M, M') \\ &= \frac{\sum_{m \in M} c(m) \times \max_{m' \in M'} \text{sim}(m, m') + \sum_{m' \in M'} c(m') \times \max_{m \in M} \text{sim}(m, m')}{|M| + |M'|} \end{aligned} \quad (7)$$

where  $c(m)$  is the frequency of  $m$  in the descriptor set  $M$  and  $c(m')$  is the frequency of  $m'$  in the descriptor set  $M'$ .

The change of citation of a cited article  $ca$  at period  $t$  is determined by the similarity of  $ca$ 's citations at  $t$  and  $t - 1$ . It is computed as

$$\text{Change}_{\text{topic}}^t(ca) = 1 - \text{sim}(D^{t-1}, D^t) \quad (8)$$

where  $D^t = \{d | d \text{ cited } ca \text{ and published at } t\}$ .

## 4. RESULTS

In this section, we compared three ways of constructing citation contexts to identify the possibly best practice for representation learning. Based on the choice of constructing citation contexts, we preliminarily investigated the characteristics and the patterns of citation context changes by applied our proposed metric on PMC OAS dataset. At last, we conducted two simple applications to show the practical potential of our proposed representation learning method and metric.

### 4.1. Data Description

We used full-text scientific articles from PMC OAS in a recent decade for our analysis and divided the decade into five periods for further analysis. The articles without a citation of identifiable articles in the full text were excluded. 1,205,407 publications have at least one effective citing sentence, and they have 31 citing sentences on average. In recent 6 years, much more articles with references have been available in PMC OAS. Each cited article roughly received 3 in-text citations on average within each period. In this study, we aim to represent cited articles by their citation contexts, so we focus on the cited articles (CA) which have enough citation context information for representation learning. We identified cited articles with more 50 in-text citations for further analysis. We show the data descriptions in Table 1.

### 4.2. Comparison Results

To compute the change score of a cited article  $ca$  at  $t$ , both representation of  $ca$  at  $t - 1$  and  $t$  would be used. For each  $t$ , only  $ca$  has more than 50 in-text citations at both  $t - 1$  and  $t$  were used in this evaluation for the robustness. We used 11,628 cited articles within the five periods for evaluation.

We used Spearman's rank correlation and Kendall's tau correlation analysis for evaluation. The correlation analysis allow comparing the similarity of ordered two types of changes scores. The results of correlation analysis is shown in Table 2. The results of two analysis are consistent. The change scores derived from three types of citation context are significantly correlated with the topical change score. The results of **WITH\_CITATION** have highest correlation coefficients. **WITH\_CITATION** can produce citation representations reflecting topical changes best. Therefore, we used citation representations derived from **WITH\_CITATION** for further investigation.

### 4.3. Distribution of Change Scores

We computed the change score for cited articles at each period and observed the average of the scores by five groups (see

**TABLE 1** | Publications and cited publications in PMC OAS from 2007 to 2016.

Period	Articles	Articles with references	Cited articles	Times cited	Cited articles <sup>a</sup> with citations > 50
2007–2008	69,394	49,105	1,431,012	3,138,614	425
2009–2010	138,204	92,773	2,523,797	6,265,277	1,568
2011–2012	252,714	225,101	4,283,203	12,288,738	4,841
2013–2014	398,620	360,902	6,367,475	19,653,641	8,994
2015–2016	502,523	477,526	8,019,222	24,905,208	11,809

<sup>a</sup>In-text citation times of cited articles.

**TABLE 2** | Comparing citation contexts.

	ONLY_CITATION		WITH_CITATION		FULL_SET	
	Coeff.	P-value	Coeff.	P-value	Coeff.	P-value
Spearman's rank	0.082	$9.5 \times 10^{-19}$	<b>0.305</b>	$2.1 \times 10^{-249}$	0.285	$7.0 \times 10^{-217}$
Kendall's tau	0.055	$1.2 \times 10^{-18}$	<b>0.208</b>	$3.3 \times 10^{-247}$	0.194	$1.1 \times 10^{-216}$

The highest coefficients are in bold.

**TABLE 3** | Temporal distribution of the change scores of cited articles from 2007 to 2016.

Period	50 < citations ≤ 56	56 < citations ≤ 64	64 < citations ≤ 77	77 < citations ≤ 107	107 < citations
2007–2008	0.130 (SD = 0.045, N = 74)	0.132 (SD = 0.047, N = 69)	0.132 (SD = 0.044, N = 68)	0.125 (SD = 0.045, N = 78)	0.123 (SD = 0.050, N = 63)
2009–2010	0.142 (SD = 0.058, N = 263)	0.136 (SD = 0.053, N = 227)	0.127 (SD = 0.043, N = 262)	0.132 (SD = 0.048, N = 280)	0.119 (SD = 0.049, N = 239)
2011–2012	0.124 (SD = 0.045, N = 956)	0.120 (SD = 0.043, N = 860)	0.119 (SD = 0.042, N = 839)	0.112 (SD = 0.038, N = 840)	0.104 (SD = 0.037, N = 867)
2013–2014	0.116 (SD = 0.043, N = 1,706)	0.113 (SD = 0.041, N = 1,611)	0.110 (SD = 0.040, N = 1,583)	0.106 (SD = 0.039, N = 1,700)	0.095 (SD = 0.035, N = 1,700)
2015–2016	0.108 (SD = 0.042, N = 2,234)	0.104 (SD = 0.039, N = 1,994)	0.102 (SD = 0.037, N = 2,168)	0.098 (SD = 0.035, N = 2,223)	0.086 (SD = 0.033, N = 2,379)

SD, Standard deviation; N, the Number of observed cited articles.

**Table 3).** These five groups divided the cited articles from 2007 to 2016 into groups with roughly even number of cited articles. Each group has about 20% cited articles over each period and the cited times of the articles within a group are in the same interval. We use the groups to observe the differences of citation changes over time and citation counts.

The change scores differ slightly between groups. The change scores of the first four groups are relatively consistent, but most of the periods have lower change scores in the fifth group. The lower change scores in the fifth group may be caused by high citations of cited articles in this group, which may indicate that highly cited articles are relatively stable in terms of their roles in science. It is reasonable to expect that the scientific community has a more rigid consensus on the scientific contribution of a more highly cited article.

The change scores show a slightly decreasing trend over time within each group, but its underlying factors remain unclear. A possible factor is the change of PMC OAS's journal coverage, because the journal coverage has effects on the completeness of semantic information for representation learning. However, the effects of the coverage need to be validated and proved by further evidence.

We also analyzed change score value distribution of the five groups (see **Figure 4**). The group of more frequently cited articles is encoded by more intensely orange. The five groups have similar distributions where most of the cited articles' change scores lie in the range of 0.04 to 0.2 and peak in the range of 0.06 to 0.1. Additionally, the distributions are roughly normal, but the groups with more than 107 citations are less slightly peaked in a lower score than the ones with fewer citations. It is consistent with our observations in **Table 3**.

## 4.4. Applications

We show two simple application examples by using the change scores and temporal citation representations to identify and understand the citation changes of cited articles.

### 4.4.1. Identifying Cited Articles With Most Changing Citation Contexts

We listed 5 cited articles with highest average change scores over recent 5 years (2012–2016) in **Table 4**. These articles have greatly changed descriptions within the full text of articles cited these articles in recent 5 years. The high change scores

may be indications of various reasons, such as high novelty or controversy. The underlying reasons need a further examination.

#### 4.4.2. Understanding the Changing Citation Contexts

The change scores alone are not informative for us to understand the changes of citation contexts of an article. Based on the citation representations, we can retrieve a series of similar words and articles at each time to interpret the changes. We use the article with the highest average change score over recent 5 years (Olsen et al., 2006) as an example to demonstrate the interpretability of the temporal representations of publications. We listed the most similar articles and a group of most similar words of the publication at each year from 2012 to 2016 in **Table 5**. We can see the words which describe the original content of the article like “phosphosite”; we can also see the words describing the scientific development related to this publication like “kinase specific phosphorylation site prediction” and “UbPred.” It is worth noting that most similar items are other articles rather than words. It is quite reasonable because publications naturally share more syntactic and semantic features than with words.

**TABLE 4 |** Top 5 publications with highest average change score over recent 5 years (2012–2016).

PubMed ID	Published year	In-text citations	Actual citations <sup>a</sup>	Avg. change score
17081983	2006	392	2,804	0.186
18171944	2003	423	1,656	0.156
19372393	2005	379	2,194	0.151
12845331	2008	594	1,684	0.143
19608861	2009	375	1,711	0.141

<sup>a</sup>From Google Scholar.

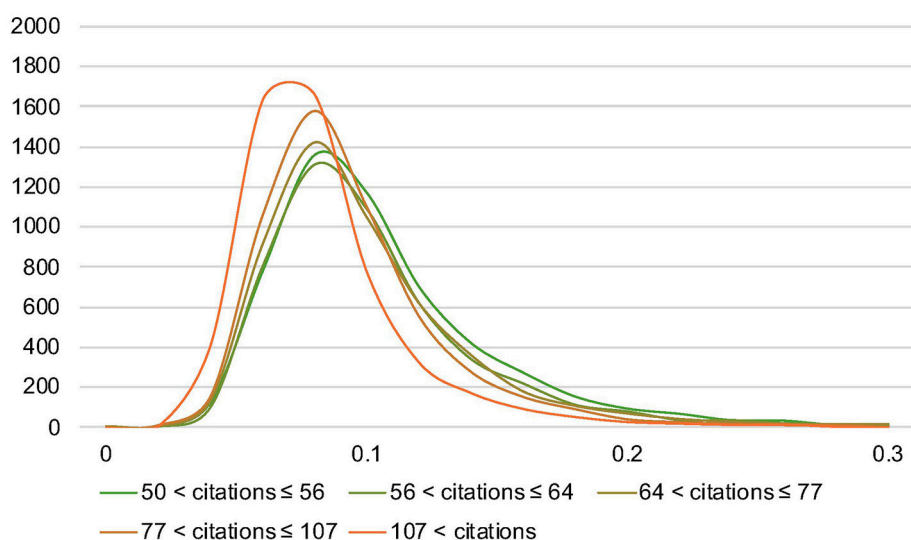
The similar articles may also provide a proxy to understand the changes.

## 5. DISCUSSION

We compared different types of citation context for learning citation representations and offered methods for identifying and understanding the changing roles of cited articles played in scientific dynamics.

We quantified the change rate of citation contexts by citation embeddings and analyzed the distribution of changes scores of a large set of biomedical articles. Both of the average and standard deviations of change scores over article groups differ in a small range. Besides, article groups have a similar distribution of change scores. These observations indicate the stability of the metric at the group level. Meanwhile, from the normal distributions in **Figure 4**, we observed a significant individual variability that can distinguish cited individual articles greatly. The change score we proposed is not only stable but also effective for identifying outstanding individuals.

Citation is a fundamental feature of scholarly communication used by researchers to position their research and lend support for claims they made (Mansourizadeh and Ahmad, 2011). Citation has become a well-established proxy for measuring scholarly impact (Garfield, 1979). Various citation-based techniques have been developed and applied to delineating and analyzing scientific structures and dynamics (Kessler, 1963; Small, 1973). Although cited articles play a dynamic role in the development of science, the dynamic aspect of citations characterized by the full text of articles hasn't been emphasized in citation-based techniques due to the lack computational and interpretable citation representation. The methods proposed for representing citations and the metric for quantifying the changes



**FIGURE 4 |** The distribution of change score value over groups. This figure is a revised version from a workshop paper that was presented at CLBib-2017 (He and Chen, 2017).

**TABLE 5 |** The changes of citation contexts of Olsen et al. (2006).

Year	Change score	The most similar publications <sup>a,b</sup>	Most similar publications <sup>b</sup>
2012	0.168	20068231 (0.84)	PHOSIDA (0.78), PhosPhAt (0.77), PhosphoSite (0.76), MiCroKit (0.75), NetPhos (0.74), ChloroP1.1 (0.74), CisGenome (0.74), Phospho.ELM (0.74)
2013	0.215	21177495 (0.79)	Guittard (0.74), Scansite(0.70), phosphosite, (0.70), Tyr216 (0.69), pY (0.68), AKT (0.68), Sarbassov(0.68), IRAG (0.68), phospho-protein (0.68)
2014	0.198	21183079 (0.83)	phosphopeptide (0.73), Phosida (0.73), NetPhos (0.73), ChIP-seq (0.73), SignalP4.1 (0.72), Scansite (0.72), kinase specific phosphorylation site prediction (0.72), mNgn2 (0.71)
2015	0.145	21183079 (0.83)	phosphosite (0.75), phosphopantetheinylation (0.71), OGlcNAcylation, (0.71), Schwannhauser (0.70), phosphotyrosine-containing (0.70), phosphopeptide (0.69), phosphoamino (0.69), PTM (0.69)
2016	0.201	21081558 (0.84)	Hornbeck,(0.80), UbPred (0.79), PHOSIDA (0.79), NetPhosK (0.78), PhosphoSitePlus, (0.78), NetPhos (0.76), PhosphoSite (0.76), pY (0.75)

<sup>a</sup>PMID. <sup>b</sup>The value in the parentheses is similarity score.

have a variety of practical implications for improving the citation-based techniques at the individual and group levels.

The representation and metric can reveal important dynamics of individual articles in the evolution of science. First, the metric has potential to identify articles of importance or interests in many applications, such as academic article recommendation and information retrieval. Second, the impact dynamics of an article may be interpreted by the metrics and the representations, for example, understanding the sudden attention attracted by a *sleeping beauty* (Van Raan, 2004) in science and identifying underlying changes of an article's impact. Third, the metric may have the ability of serving as an early indication of an article's impact dynamics. Forth, the metric may provide supplementary information for scientific evaluation based on citation.

The representation and metric may also be used to enhance citation-based approaches to science mapping, such as bibliographic coupling (Kessler, 1963) and co-citation analysis (Small, 1973). Integrating the metric with citation-based approaches can reveal scientific dynamics that conveys foresights into emerging trends (Chen, 2016). For example, a cluster of articles where many of the articles cited references in new contexts may be an early sign of an emerging research topic.

## REFERENCES

Berger, M., McDonough, K., and Seversky, L. M. (2017). Cite2vec: citation-driven document exploration via word embeddings. *IEEE Trans. Visual. Comput. Graph.* 23, 691–700. doi: 10.1109/TVCG.2016.2598667

## 6. CONCLUSION

This study has limitations and we plan to improve our methods and further investigate the factors affecting the change of citation context. We didn't use the information of how a citation was mentioned in a sentence in the representation learning. For example, a citation can play an explicit grammatical role within a sentence or play no explicit grammatical role in a sentence usually by being placed within a bracket (Thompson and Tribble, 2001). In the future, we will construct different contexts for citations with different forms. The investigation on the factors affecting the changes of citation contexts in this study is limited. We will investigate more factors and their effects. The mechanism of how the changes of citation contexts affect future impact of cited articles is another interesting question we will study in the future.

In conclusion, we introduced an embedding learning method to represent scientific articles by using the citation context text in other articles. Our method emphasizes the temporal features of citation text to characterize the dynamic role of scientific publications. The temporal representation can be used to quantify how much the role of a publication changed as well as interpret how the role changed over time. Base on the study on a large biomedical full-text literature dataset, we evaluated different citation contexts for representing citation over time and found that using sentences with in-text citation reflect topical change best. We also concluded that the metric for quantifying the changes of articles' roles is stable over time at the population level and there is significant individual variability to distinguish individuals. We hope these insights will facilitate further research into improving citation-based indicators and analysis approaches by modeling citation contexts.

## AUTHOR CONTRIBUTIONS

JH designed the study, conducted the experiment, and wrote the first version of the manuscript. CC designed the evaluation method, improved the methods of the study and revised the manuscript.

## FUNDING

The work is supported by the National Science Foundation (Award Number: 1633286).

## ACKNOWLEDGMENTS

This is an extended and revised version of a workshop paper that was presented at CLBib-2017 (He and Chen, 2017). We thank Zhipeng Zheng at Department of Chemistry, University of Pennsylvania for his help in the example interpretation.

Boyack, K. W., Small, H., and Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *J. Assoc. Inform. Sci. Technol.* 64, 1759–1767. doi: 10.1002/asi.22896

Callahan, A., Hockema, S., and Eysenbach, G. (2010). Contextual cocitation: augmenting cocitation analysis and its applications. *J. Assoc. Inform. Sci. Technol.* 61, 1130–1143. doi: 10.1002/asi.21313



- Chen, C. (2012). Predictive effects of structural variation on citation counts. *J. Am. Soc. Inform. Sci. Technol.* 63, 431–449. doi: 10.1002/asi.21694
- Chen, C. (2016). Grand challenges in measuring and characterizing scholarly impact. *Front. Res. Metr. Anal.* 1:4. doi: 10.3389/frma.2016.00004
- Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., and Radev, D. (2008). Blind men and elephants: what do citation summaries tell us about a research article? *J. Assoc. Inform. Sci. Technol.* 59, 51–62. doi: 10.1002/asi.20707
- Eto, M. (2013). Evaluations of context-based co-citation searching. *Scientometrics* 94, 651–673. doi: 10.1007/s11192-012-0756-z
- Ganesh, J., Ganguly, S., Gupta, M., Varma, V., and Pudi, V. (2016). “Author2vec: Learning author representations by combining content and link information,” in *Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion*. (Geneva: International World Wide Web Conferences Steering Committee), 49–50.
- Ganguly, S., and Pudi, V. (2017). “Paper2vec: combining graph and text information for scientific paper representation,” in *European Conference on Information Retrieval* (Aberdeen: Springer), 383–395.
- Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics* 1, 359–375. doi: 10.1007/BF02019306
- Gipp, B., and Beel, J. (2009). “Citation proximity analysis (CPA): a new approach for identifying related work based on co-citation analysis,” in *ISSI-09: 12th International Conference on Scientometrics and Informetrics*, (Rio de Janeiro) 571–575.
- Gipp, B., Meuschke, N., and Lipinski, M. (2015). “CITREC: an evaluation framework for citation-based similarity measures based on TREC genomics and pubmed central,” in *Proceedings of the iConference 2015* (Newport Beach, CA).
- Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). “Diachronic word embeddings reveal statistical laws of semantic change,” in *Proceedings Association Computational Linguistics* (Berlin: ACL).
- He, J., and Chen, C. (2017). “Understanding the changing roles of scientific publications via citation embeddings,” in *Proceedings of the Second Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics (CLBib-2017) Co-located with 16th International Conference on Scientometrics and Informetrics (ISSI 2017)* (Wuhan), 42–48.
- He, J., and Chen, C. (2018). Predictive effects of novelty measured by temporal embeddings on growth in science. *Front. Res. Metr. Anal.* 3:9. doi: 10.3389/frma.2018.00009
- He, Q., Pei, J., Kifer, D., Mitra, P., and Giles, L. (2010). “Context-aware citation recommendation,” in *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, NC: ACM), 421–430.
- Huang, W., Wu, Z., Liang, C., Mitra, P., and Giles, C. L. (2015). “A neural probabilistic model for context based citation recommendation,” in *AAAI 2015: Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence*, 2404–2410.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *J. Assoc. Inform. Sci. Technol.* 14, 10–25. doi: 10.1002/asi.5090140103
- Lin, D. (1998). “An information-theoretic definition of similarity,” in *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)* (San Francisco, CA) 296–304.
- Liu, S., and Chen, C. (2012). The proximity of co-citation. *Scientometrics* 91, 495–511. doi: 10.1007/s11192-011-0575-7
- Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., and Lin, Y. (2014a). Literature retrieval based on citation context. *Scientometrics* 101, 1293–1307. doi: 10.1007/s11192-014-1233-7
- Liu, X., Yu, Y., Guo, C., Sun, Y., and Gao, L. (2014b). “Full-text based context-rich heterogeneous network mining approach for citation recommendation,” in *IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on IEEE* (London), 361–370.
- Mansourizadeh, K. and Ahmad, U. K. (2011). Citation practices among non-native expert and novice scientific writers. *J. Engl. Acad. Purp.* 10, 152–161. doi: 10.1016/j.jeap.2011.03.004
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). “Distributed representations of words and phrases and their compositionality,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13* (Curran Associates Inc.), 3111–3119.
- Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishnan, P., Qazvinian, V., et al. (2009). “Using citations to generate surveys of scientific paradigms,” in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics) (Boulder, CO), 584–592.
- Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., et al. (2006). Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* 127, 635–648. doi: 10.1016/j.cell.2006.09.026
- Orosz, K., Farkas, I. J., and Pollner, P. (2016). Quantifying the changing role of past publications. *Scientometrics* 108, 829–853. doi: 10.1007/s11192-016-1971-9
- Qazvinian, V., Radev, D. R., and Özgür, A. (2010). “Citation summarization through keyphrase extraction,” in *Proceedings of the 23rd International Conference on Computational Linguistics, COLING '10*. (Stroudsburg, PA: Association for Computational Linguistics), 895–903.
- Řehůřek, R. and Sojka, P. (2010). “Software framework for topic modelling with large corpora. in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (Valletta: ELRA), 45–50.
- Resnik, P. (1995). “Using information content to evaluate semantic similarity in a taxonomy,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'95* (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 448–453.
- Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. Assoc. Inform. Sci. Technol.* 24, 265–269.
- Small, H. (2011). Interpreting maps of science using citation context sentiments: a preliminary investigation. *Scientometrics* 87, 373–388. doi: 10.1007/s11192-011-0349-2
- Small, H., and Klavans, R. (2011). “Identifying scientific breakthroughs by combining co-citation analysis and citation context,” in *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics* (Leuven), 783–793.
- Sugiyama, K., and Kan, M.-Y. (2013). “Exploiting potential citation papers in scholarly paper recommendation,” in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries* (New York, NY: ACM), 153–162.
- Teufel, S., Siddharthan, A., and Tidhar, D. (2006). “Automatic classification of citation function,” in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (Sydney, NSW: Association for Computational Linguistics), 103–110.
- Thompson, P., and Tribble, C. (2001). Looking at citations: using Corpora in English for academic purposes. *Lang. Learn. Technol.* 5, 91–105.
- Tian, H., and Zhuo, H. H. (2017). Paper2vec: citation-context based document distributed representation for scholar recommendation. *arXiv [preprint] arXiv:1703.06587*.
- Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics* 59, 467–472. doi: 10.1023/B:SCIE.0000018543.82441.f1
- Zhu, S., Zeng, J., and Mamitsuka, H. (2009). Enhancing medline document clustering by incorporating mesh semantic similarity. *Bioinformatics* 25, 1944–1951. doi: 10.1093/bioinformatics/btp338

**Conflict of Interest Statement:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 He and Chen. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.