A Proposed Method for Residual Citation Allocation Based on Citation Contexts' Similarity

Toluwase Victor Asubiaro (tasubiar@uwo.ca)
University of Western Ontario https://orcid.org/0000-0003-0718-7739
Isola Ajiferuke
University of Western Ontario

Research Article

Keywords: residual citations, scholarly communication impact, citation analysis, citation context analysis, semantic similarity

Posted Date: December 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1041491/v1

Abstract

This article proposes an approach for allocating residual citations to scientific publications and demonstrating this proposed approach with a sample of biomedical publications. Residue citations (i.e., citations that are lost due to citation practices termed “Obliteration by Incorporation” and the “Palimpsestic Syndrome”) in consequent citations in the second, third or nth generations are then reconstituted. The proposed approach takes into account citation contexts (i.e., the contribution of a cited publication) for allocating residual citation. The proposed method for allocating residual citation is based on the similarity between the citation contexts of a publication and those of its nth generation citations in their n+1th generation citations. The proposed method was demonstrated using a sample with ten base articles and their five generations of citations, from which 5,272 citation context pairs were obtained. The proposed indirect citation weighting was compared with the existing cascading citation weighting method using one T-test. Statistical tests were also performed to understand the differences in the residual citations from one generation to the other. Like the cascading citation system, residual citations from articles to their generations of citations decreased as the number of generations increased. However, residue citations accrued to publications at all the generations were statistically different between the proposed residual citation and the cascading citation system. This study proposes a method for assessing scientific communication based on the contribution of scientific publications beyond the conventional direct citation.

Introduction

Citation is an integral part of the scientific culture and ecosystem. Citation analysis, and bibliometrics in general, as a way of studying science, is science (McKeown et al., 2016). Citations, as records and art (of citing), are considered sacrosanct in science as citations are made in and produced from a seemingly reputable scholarship system where researchers are mandated to make references to original owners of the ideas that they have consulted, referenced or used in scholarly communications. Citations in science refer explicitly to the credits or references made to intellectual contributions published in journal articles, books or book chapters, web pages, conference proceedings, and other scholarly communication channels by users of the published information. Citations are made to attribute ideas, opinions, results or observations that were made in prior research to the source (Wan & Liu, 2014).

Quality assessment in the research community today is primarily built around citations. Though scientific publications go through the critical peer-review process, citations have more influence in the evaluation of science. According to MacRoberts and MacRoberts (1989) and Wallin (2005), very highly cited papers over time are considered more factual and are accepted as part of the universe of knowledge. The most famous metrics for assessing research are based on the simple citation count, including the journal impact factor (JIF), The EigenFactor (EF), Article influence score (AIS), P-rank, SNIP, PageRank metric, etc.

While citation is the most popular system of assessing the contribution of a publication to science by counting the number of times it has been cited, this method has been criticized. Ordinary citation count

does not take into account the contribution of the cited document, which makes it fall short of its purpose of assessing the contribution of cited publication in the citing publication. Also, the simple citation count-based evaluation method assumes that all citations are equal. This does not reflect a real-life research evaluation task, which is a complex and multi-dimensional problem. Several studies on citation classification and weighting have shown that some citations contribute more to citing articles than others. These studies have classified or weighted citations based on functions, impact, importance, utility, location, frequency and sentiment. (Strotmann and Zhao, 2014; Wan and Liu, 2014; Zhao and Strotmann, 2016) by analyzing citation contexts.

Citation Context Paradigm

Citation context analysis incorporates the examination of citation contexts (texts around the citation marker) into citation analysis and has been found useful in addressing some shortcomings of the traditional citation analysis. Citation context is usually the text that surrounds the in-text citation "used to refer to other scientific works" (Doslu & Bingol, 2016, p. 654) and represents the context in which a citation is referenced. Citation context analysis has profoundly influenced the citation analysis field (Ding et al., 2014; Jeong et al., 2014) as they can be used to "identify the main contributions of a scientific publication" (Doslu & Bingol, 2016, p. 654). The definition of a citation context was described as the "citation's context within the full text of the scientific paper" that cited it, rather than the simple citation count (Ding et al., 2014, p. 1821). A citation context in the citing article is incomplete without considering all the mentions of the citation in the full text of the article. Citation context, therefore, could be a sentence or phrase. In some cases, a citation context can span sentences or a paragraph. Studies have suggested various windows of text to capture the extent of the contribution of the cited publication in the citing publications for citation context identification. In the review of the citation context window size by Iqbal et al., (2021), it was revealed that a range of one to four sentence-window and 50 to 100 word-window were recommended in the previous studies for citation context identification.

Citation context analysis studies have resulted in two classes of research-citation context weighting and classification. Weights are finitely or infinitely continuous quantitative values, where classes are finite categories. In some instances, weighted citations are converted to categories and vice-versa. For instance, citation context sentiment analysis have resulted in citation polarity or citation sentiments categories. While citation polarity is a continuous variable with values between +1 and -1, citation sentiment is the classification of citation polarities, where values that are greater than zero are classified as positive sentiment citations, polarities that are less than zero are classified as negative sentiment citation and polarities that are equal to zero are classified as neutral sentiment citations.

With all the developments in the citation context analysis, credit attribution beyond direct citation remains unexplored. That is, citation context and citation analysis only consider the attribution of credit between a document and the direct citations it has received. This study explores the use of citation contexts for attributing credits to scientific papers beyond the credits from documents that directly cited them. We propose a method of credit attribution or citation weighting called residual citation allocation that is

based on the contribution of a scientific publication to the paper that did not directly cite it. Citations or credit attributions received by a scientific paper from documents that did not cite it is called a residual citation.

Residual Citations on Citation Path

One of the developments in citation analysis is the cascading citation research (Dervos & Kalkanis, 2005) that proposes that scientific publications assert more influence beyond the direct citations they get. Therefore, scientific publications should receive credit as residual or indirect citations from papers citing their citations. The justification for this idea is that scientific publications are an embodiment of contributions from different publications they cited. Hence, papers that cite scientific publications benefit from the contributions from the scientific publications that were cited. The documents that cite the citations of a scientific publication are its second-generation citations, and indirect citations should be accrued from the second-generation (and subsequently nth generation) citations (Dervos & Kalkanis, 2005).

Looking at scientific publications from the perspective of a network, in practice, a publication is part of a network of publications with an established level of relationship. On research networks, two publications are either connected by citation, co-citation or bibliographic coupling, the three defining factors for relatedness. Citation paths refer to generations of citations from a previously cited publication. For instance, if there are three publications A, B, and C, where B cited A, and C cited B, there is a citation path from A to C while A-B-C for a citation chain. The three existing forms of relatedness on citation networks are only limited to A-B and B-C, with no provision for exploring the relationship between A and C. Citation from B to A or citation path from A to B is the conventional direct citation relationship. We describe the citation from C to A as indirect or residual citation.

The residual citation idea brings into play the exploration of the A-C (and beyond) relatedness. While Dervos and Kalkanis (2005) and Fragkiadaki, Evangelidis, Samaras, and Dervos (2009) pioneered the residual citation allocation idea, the proposed implementation based on the cascading citation system that recommended allocating some fractional value ($1/2^{n-1}$) for nth generation citations comes short of exploring citation context information for weighting. We also fault the cascading citation idea because it assumes that all scientific publications deserve a residual citation each time a document cites their citations. With the developments in the citation context analysis research area, it is possible to quantify the level of relatedness between publication A and publication C based on the contribution of publication A in publication C.

Justifications for allocating Residual Citation

Though we agree with the mantra of allocating residual citations to scientific publications, we propose that residual citations should be proportional to the contribution of a scientific publication in its indirect citation. We explore practical cases where residual citation allocation is a necessity in the evaluation of science.

Science does not always attribute credits to the proper sources because of some epistemic practices. These practices have resulted in the under citations of some scientific publications. McCain, (2014) identified two sources of undercitations called “Obliteration by Incorporation” and the “Palimpsestic Syndrome” as described in (Merton, 1965, 1988). “Obliteration by Incorporation” occurs in the form of “tacit citation” where the ideas, methods, or findings are “anonymously incorporated into current canonical knowledge” because of the obliteration of their sources (Merton, 1988, p. 622).

The Palimpsestic Syndrome occurs when an author attributes credit about an idea from earlier work to a more recent author in whose work the idea was first encountered (Merton, 1965). The original work losses citations in the form of “tacit citation” to more recent work. For instance, Publication C cited publication B, but the credit attributed to publication C should have been attributed to Publication A, because Publication B cited the idea from Publication A. Maybe the authors of Publication C cited Publication B because they did not have access to Publication A directly while writing their manuscript or did not allocate enough time to finding Publication C. In this instance, the influence or contribution of Publication A is rightly beyond publication B that cited it directly. With the proper allocation of residual citations, it is possible to accrue citations to the original work.

Objectives of the Study and Research Questions

This study’s aim is to propose a system for allocating residual citations based on the citation contexts’ semantic similarity and explore the pattern of residual citations from sampled scientific publications and their five generations of citations.

The following research questions are intended to guide this study:

1. How different is the proposed semantic similarity-based residual citation weights from the cascading citation weights?
2. What differences exist in the highest semantic similarity scores between the generations of citation?
3. What is the pattern of residual citations between the sampled cited documents and the nth generation citations?

Theoretical Framework

Allocating weights to indirect citation mentions is different from allocating weights to direct citations. The theoretical framework for allocating residual citation weight on a citation chain is based on the semantic similarity between the contributions of a publication in its citing article and the contribution of its nth generation citation in the n+1th generation article. Citation contexts are surrogates of cited publications' contributions.

On a citation network, citation paths or citation chains extend beyond conventional direct citations. Direct citations are only the origins of the chains, with the base article at the summit and direct citation is the node next to the base article. Using publications A, B, C, D, and E as examples of articles on a citation chain, where publication B cited publication A, publication C cited publication B; publication D cited

publication C and publication E cited publication D. Therefore, there is a citation path or citation chain A-B-C-D-E, where any of the publications, except E can be the base article. In this framework section, article A will be taken as the base article. Taken that publication A is the base article, publication B, C, D, and E is the first generation or direct citation, second generation, third generation and fourth generation citations, respectively.

In theory, residual citation does not have to accrue to the base article all the time. For example, paper B may copy a methodological section from paper A citing a source and thus citing paper A, and paper C may also copy the source methodology aspect from paper B and thus cite paper B, but does not cite paper A. In this case, publication A deserves residual citation from paper C. However, it is possible for paper C to cite an aspect of paper B (e.g. sampling methods which is completely different from the aspect that paper B cited in paper A (e.g. data analysis technique). In such a case, paper A does not deserve a residual citation from paper C.

To find the residual citation can be accrued from the second generation to paper A, for instance, we obtain the semantic similarity score between the contribution of the base article in publication B, and the contribution of publication B in publication C. Using Figure 1, it can be observed that while there are six citation contexts of paper A in paper B (A-B0, A-B1, A-B2, A-B3, A-B4, and A-B5), but only one citation context of paper B in paper C (B-C0). We need to compare all the contributions of paper A in paper B and all the contributions of paper B in paper C, thereby obtaining the following six pairs: A-B0|B-C0, A-B1|B-C0, A-B2|B-C0, A-B3|B-C0, A-B4|B-C0, and A-B5|B-C0.

Importantly, the theoretical framework assumes that the citation context pairs with the highest semantic similarity can give a hint if at least a residual citation should accrue to the base article. It is recognized that average, median, and minimum values of the semantic similarity scores could be considered. This study used the highest semantic similarity scores because this study is exploratory. Using the above example, the citation context pairs with the highest semantic similarity scores among the six pairs of citation contexts would be considered for allocating citation residual to publication A from its second-generation citation B. Therefore, if the semantic similarity score of the pair of citation contexts with the highest semantic similarity is significant enough to the categorized as "similar", then at least a contribution of publication A in publication B is similar to the contribution of publication B in publication C. Thus, publication A deserves residual citation from its second-generation citation.

Methodology

Sampling Articles at the First to Fifth Generations

Sampling was done in six stages: the first stage for the base articles and the last five stages for five generations of citations (one stage per generation). The ten most-cited biomedical articles that were published in the year 2014 were sampled as the base articles. For the second stage of sampling, only the most cited five papers that cited each of the base articles were sampled, provided their full texts were

available. Hence, for the ten base articles, this resulted in 50 first-generation citations. The next stage was sampling the second-generation citations, which was done by sampling the most cited two citations of each of the 50 first-generation citations, that is a total of 100 second-generation citations. The most cited top two citations for each of the 100 second-generation articles were also sampled as third-generation articles, making a total of 200 third-generation citations after removing duplicates. This was repeated for the fourth and fifth generations to give 400 and 800 potential fourth and fifth-generation articles, respectively.

Duplicates were not included in the data that was collected, i.e., once an article is sampled, it was not sampled again in subsequent generations. For instance, Ross et al (2016)[1] and Alexander et al (2017)[2] were second-generation citations for one of the base articles, the Alexander et al (2017) also cited the Ross et al (2016), and ranked as one of its two most cited articles citing articles, a potential third-generation sample. However, Alexander (2017) was not sampled as a third-generation article because it was sampled in an earlier generation. Some other articles were also excluded based on discretion during the sampling period. For instance, articles in other disciplines, not in biomedicine e.g. Krittanawong C. et.al. (2019)[3] created generations of articles mostly in computer science, were ignored at the first or second-generation level because articles at the subsequent generations significantly deviated from biomedicine. It was intended that the sample would be core biomedical publications, and including articles from other disciplines outside biomedicine would have defeated this purpose.

Citation Data Collection

Full texts of the sampled first to fifth-generation articles were downloaded. The citation contexts of articles in the sample was collected from the full texts of articles that cited (subsequent generation). Thus, the citation contexts of an article in a generation was collected from the full text of articles in the subsequent generation that cited it. For example, the citation contexts of a base article were collected from the full text of first-generation articles that cited it. The citation contexts of a first-generation article in the sample were collected from the full text of second-generation articles in the sample that cited it.

The citation contexts were manually collected from the full text of the appropriate article. Citation contexts were manually collected by reading the citation sentence and the sentences before and after the citation sentence to understand the expense of citation contexts. The citation contexts were not dependent on pre-specified window of texts, we manually determined the citation context by understanding the context in which a cited publication was referenced. Thus, citation contexts were of different lengths, from a phrase, one sentence, to many sentences.

Citation contexts in Tables, Images, and supplementary materials were not included. Citation contexts were preprocessed by removing texts that did not add semantic value (e.g. numbers). Implicit in-text citation markers were removed while explicit in-text citation markers were replaced. Examples of implicit in-text citation marker include (Asubiaro, 2020), (Asubiaro et.al., 2020), 123 and ^{123}. Implicit in-text citation markers did not add semantic value to citation contexts and were removed. On the other hand,

examples of explicit in-text citation markers include Sergio (2020) or Sergio et. al. (2020). Explicit in-text citation markers are part of the semantic structure of the citation contexts and were replaced with in_text_ref to reduce the number of out-of-vocabulary (OOV) words in the data collected. The accuracy of textual data model increases with the reduction in the number of OOVs in textual data. Ambiguous acronyms (e.g. ER could refer to an emergency room or endoplasmic reticulum) were replaced with full meanings.

Indirect Citation Data Weighting

Citation contexts of two sets of mentions were collected and were paired up. For example, if the base article was mentioned in a first-generation citation three times, and the first generation citation was mentioned in the second generation citation four times. Let us represent these mentions as follows: M_{a_01},M_{b_01},andM_{c_01}; M_{a_12}, M_{b_12}, M_{c_12}, and M_{d_12}. By pairing each citation context from the first set against each of the citation context in the second set, we obtained the following 12 citation context pairs: M_{a_01}| M_{a_12}, M_{a_01}| M_{b_12}, M_{a_01}| M_{c_12}, M_{a_01}| M_{d_12}, M_{b_01}| M_{a_12}, M_{b_01}| M_{b_12}, M_{b_01}| M_{c_12}, M_{b_01}| M_{d_12}, M_{c_01}| M_{a_12}, M_{c_01}| M_{b_12}, M_{c_01}| M_{c_12}, and M_{c_01}| M_{d_12}. We then calculated the similarity score between the two citation contexts in each pair, and the highest similarity score is assigned as the residual citation accruing to the base paper from the second generation citation.

To illustrate the method described above, in Figure 2, a paper was cited by two first generation articles (FirstGen-article1 and FirstGen-article2), and each of the two first generation articles were cited by two second generation articles- while FirstGen-article1 was cited by SecondGen-article1 and SecondGen-article2, FirstGen-article2 was cited by SecondGen-article3 and SecondGen-article4. The base article was referenced once in the FirstGen-article1, while FirstGen-article1 was mentioned seven times in SecondGen-article1 and five times in SecondGen-article2. On the other hand, the base article was mentioned six times in FirstGen-article2 while both SecondGen-article3 and SecondGen-article4 referenced FirstGen-article2 once. The citation context pairs with the highest semantic similarity measures were indicated with arrows and marked 1 to 4.

Citation Context Similarity Measure

Manually collected and preprocessed citation contexts pairs were inputted into a python program that returned the semantic similarity scores between the citation context pairs. The python script implements the BioSentVec sentence embeddings models, a state-of-the-art for biomedical scientific publications language representation, trained on PubMed dataset with over 28 million biomedical scientific publications (Chen et al., 2019).

The python script generates sentence vectors given any arbitrary sentences as inputs, and the cosine of the angle between the representations of the sentences on vector space is the semantic similarity between the two sentences.

Citation Context Pair Classification and Weight Allocation

Citation context pairs were classified into three classes ("similar", "somewhat similar", and "not similar") for weights allocation. Semantic similarity scores between citation context pairs from the Python script is a value between zero and one. Where citation similarity scores close to 1 indicates citation context pairs are similar while semantic similarity scores close to zero indicates citation context pairs are not similar. This study adopted the thresholds for the three classes of semantic similarities of citation contexts in Asubiaro, (2021). (Asubiaro, 2021) used expert annotated citation context pairs from biomedical publications to train computer algorithms that classified citation context pairs with < 0.51 semantic similarity scores as not similar; citation context pairs with semantic similarity scores > or =0.51 and <0.71 as somewhat similar and citation context pairs with semantic similarity score >or =0.71 as similar.

Similar Citation context pairs were allocated a weight of 0, somewhat similar citation context pairs were allocated a weight of 0.5 and “not similar” citation context pairs were allocated a weight of 1.

Results

The number of articles sampled at every generation of citation, as well as the number of citation contexts at every citation generation for each of the ten base articles, can be found in Table 1. Included in Table 2 are the number of citation contexts collected from generations one to five. The average number of citation contexts of the base articles that were extracted from the first generation was 9.8 (maximum=20, minimum=5). A total of 221 (average=22.1, maximum=44, minimum=13) citation contexts of the first-generation articles were extracted from the second-generation articles. 439 (average=43.9, maximum=61, minimum=22) citation contexts of the second-generation articles were extracted from the third-generation articles. Fourth-generation articles produced 748 (average=74.8, maximum=102, minimum=40) citation contexts of third-generation articles. Similarly, fifth-generation articles produced 1257 (average=125.7, maximum=141, minimum=113) citation contexts of fifth-generation articles. For ease of reporting, citation contexts of the base articles from the first-generation articles were labelled first-generation citation contexts. Similarly, citation contexts of the first-generation articles that were obtained from the second-generation articles were labelled second-generation citation contexts. The same rule applies to the citation from other generations.

The number of citation context pairs between the citation contexts of first-generation articles and citation contexts of articles in other generations is displayed in Table 4.6. The result shows that the number of citation context pairs from the first- and second-generation citation contexts was 419 (average= 41.9, maximum=103, minimum=15). The number of citation context pairs from the first- and third-generation citation contexts was 879 (average=87.9, maximum=194, minimum=36). The number of citation context pairs from the first- and fourth-generation publications was 1524 (average=152.4, maximum=259, minimum=79). The number of citation context pairs from the first- and fifth-generation publications is 2450 (average=245.0, maximum=563, minimum=127).

The number of citation pairs depends on the number of citation mentions in the citing articles of the two generations in question. The number of citation context pairs between an article with m citation mentions

and another article with ***n*** citation mentions was obtained as ***n*x*m***. The lowest citation context pairs (n=249) from the first and other generation papers were recorded by the second article, while the highest number (n=1,114) of citation context pairs were from the ninth article. The total number of citation context pairs for the indirect citation weighting part of this thesis was 5272.

Table 1: Indirect Citations Statistics

	Table	times cited	sample size per generation					citation contexts no/generation				
			1	2	3	4	5	1	2	3	4	5
1	Siegel, R. et al (2014) Cancer statistics, 2014.	9090	5	10	20	40	69	11	22	22	40	113
2	Bolger, A.M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data.	8864	5	10	20	40	74	5	15	36	71	127
3	Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.	8508	5	10	20	40	73	8	19	41	76	128
4	Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.	6732	5	10	20	40	78	6	17	41	65	131
5	Ogden, C.I. et al (2014). Prevalence of childhood and adult obesity in the United States, 2011-2012.	4928	5	10	20	40	67	10	15	60	89	112
6	Ng., M. (2014). Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013.	4576	5	10	20	40	68	10	44	61	83	130
7	Go, A., et al (2014). Photovoltaics. Interface engineering of highly efficient perovskite solar cells.	3905	5	10	20	40	67	7	13	52	76	141
8	Koln, P., et al (2014). Heart disease and stroke statistics--2014 update: a report from the American Heart Association.	3880	5	10	20	40	65	12	21	42	102	117
9	Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) Black phosphorus field-effect transistors.	3577	5	10	20	40	75	20	30	51	68	137
10	Lamouille, S., Xu, J., and Derynck, R. (2014). Solvent engineering for high-performance inorganic-organic hybrid perovskite solar cells.	3323	5	10	20	40	74	9	25	33	78	121
			50	100	200	400	710	98	221	439	748	1257

Table 2: Number of citation context pairs in the Indirect Citation Dataset

	No of pairs between first and other generations				
	First-second	First-third	First-fourth	First-fifth	Total
Article 1	34	89	168	266	557
Article 2	15	36	71	127	249
Article 3	28	59	117	208	412
Article 4	21	45	79	151	296
Article 5	29	97	171	238	535
Article 6	74	122	163	221	580
Article 7	19	82	108	202	411
Article 8	55	107	251	276	689
Article 9	103	194	254	562	1113
Article 10	41	49	138	202	430
Total	419	880	1520	2450	5272

Descriptive Statistics of the semantic similarity measure between citation context pairs are presented in Table 3. below. First, the distributions of the semantic similarity measures on histogram graphs were inspected visually for normality. The distributions are presented in Figure 3, Figure 4, Figure 5 and Figure 6, and they are symmetrical in shape. Table 3 shows that averages of the weights of the residual citations received by the base articles from the second, third, fourth and fifth generations are 0.47, 0.43, 0.40 and 0.37, respectively. The average reduced consistently from the second to the fifth citation generation.

Table 3: Averages of the residual citations received from the second to the fifth citation generations

	First-second	First-third	First-fourth	First-fifth
article 1	0.51	0.41	0.40	0.39
article 2	0.31	0.33	0.30	0.28
article 3	0.44	0.41	0.36	0.31
article 4	0.3	0.27	0.29	0.27
article 5	0.5	0.44	0.43	0.40
article 6	0.52	0.5	0.45	0.42
article 7	0.41	0.41	0.40	0.38
article 8	0.62	0.57	0.53	0.48
article 9	0.59	0.5	0.48	0.42
article 10	0.48	0.45	0.41	0.40
all	0.47	0.43	0.40	.37

Indirect Citation Weights

The semantic similarity-based residual citation weights were categorized using the thresholds that were specified in the methods section for classifying the citation weights. **Not similar** citation context pairs (i.e. with less than 0.51 semantic similarity score) were allocated zero weight. **Somewhat similar** citation context pairs (i.e., greater than or equal to 0.51 and less than 0.71 semantic similarity score) were allocated a weight of 0.5. **Similar** citation context pairs (i.e. greater than or equal to 0.71 semantic similarity score) were allocated weight of one.

Categorization of the weights (Table 4) shows that the fewest of weights received by the base articles was that of 1. Most of the weights received were zero and the proportion of zero weights increased from second generation to the fifth generation.

Table 4: Categories of the residual citation semantic similarity scores

Generation	Weight=1	Weight=0.5	Weight=0	N
Second	4%	37.00%	59%	100
Third	0	26.50%	73.50%	200
Fourth	1%	20%	79%	400
Fifth	0%	10%	90%	710

The percentage of non-zero weights received by the base articles from the second to the fifth generations are presented in Table 5. The result shows the percentage of non-zero weights received by each of the base articles reduced from the second generation (43%) to the fifth generation (10%). Article 8 consistently received that highest percentage of non-zero weight at all the generations, with 90% non-zero weight at the second generation, more than 50% non-zero weights at all the generations except the 5^{th} generation. On the other hand, the worst-performing base articles-Article 2 and Article 4- received no non-zero residual weights at three of the four generations of citations.

Table 5: Non-zero indirect Citation weights

Base articles	Non-zero residual weights			
	2^{nd} Generation	3^{rd} Generation	4^{th} Generation	5^{th} Generation
article 1	40%	15%	22.5%	10.14%
article 2	0%	5%	0%	0%
article 3	10%	25%	12.5%	2.74%
article 4	0%	0%	5%	0%
article 5	30%	30%	17.5%	8.96%
article 6	80%	45%	17.5%	13.24%
article 7	20%	10%	12.5%	7.35%
article 8	90%	80%	60%	36.92%
article 9	90%	50%	37.5%	18.67%
article 10	50%	5%	12.5%	5.41%
Total	43%	26.5%	19.75%	10%

Statistical Test

From the observations in Table 4, the averages of the semantic similarity scores between the citation context pairs reduced from the first generation to the fifth. This observation was consolidated with the number of non-zero weights in Table 6 as the number of non-zero weights also reduced from the first to fifth generation citations. To find out if the differences in the averages are statistically significant, **Hypothesis 6_0** was stated and tested.

Hypothesis 6_0: The residual citation score per paper is the same for all the generations of citation.

The averages of the semantic similarity scores between the citation context in the first- generation articles and subsequent generations decreased as the generations got farther from the base article. In other words, using semantic similarity score between the citation contexts as a measure of the average knowledge flow from the base article, the result of the averages of the semantic similarity measure shows that knowledge flow from the base article continuously reduced as the generations of citations increased. Therefore, **Hypothesis 6_0** was stated to guide this thesis. Inferential statistics were therefore performed to confirm if the observed differences in the averages are significant.

The data is continuous, a recommended statistical test is the analysis of variance (ANOVA). It was tested if the datasets conformed to other conditions for ANOVA test. The following conditions were examined:

1. Dependent variable (interval data type): semantic similarity scores
2. Normally distributed samples: The histogram of the four distributions are displayed in Figure 3, Figure 4, Figure 5 and Figure 6, which shows all the distributions are approximately normal.
3. Test of Homogeneity: Result of the Levene's test of homogeneity of variances is displayed in Table 6. We reject the null hypothesis as $p<0.05$. The variances are not equal. The datasets violated the test of homogeneity of variances; therefore, the datasets are not appropriate for ANOVA. Kruskal Wallis, a non-parametric test, was considered as an alternative to ANOVA to test if the differences between the citation context pairs' semantic similarity scores are significant.

Table 6: Tests of Homogeneity of Variances for the semantic similarity score per paper

	Levene Statistic	df1	df2	Sig.
Based on Mean	5.333	3	1406	.001
Based on Median	5.125	3	1406	.002
Based on Median and with adjusted df	5.125	3	1367.793	.002
Based on trimmed mean	5.345	3	1406	.001

Kruskal-Wallis Test

The result of the Kruskal-Wallis statistic test is displayed in Table 7 below. A Kruskal-Wallis test showed there was a statistically significant difference in the semantic similarity score per paper between the generations of citation, $\chi2(3) = 65.58$, $p = 0.00$, with a mean rank semantic similarity score of 917.31 for the second generation citations, 817.79 for the third generation citations, 731.23 for the third generation citations, and 629.54 for the fifth generation citations. The mean rank statistic shows that the citation context similarities reduced as the generations went farther from the base article, and this is statistically significant.

Table 7: Mean Rank Statistics

	Semantic similarity categories	N	Mean Rank
Residual citation weights score per paper from the four generations	Second generation citations	100	917.31
	Third generation citations	200	817.79
	Fourth generation citations	400	731.23
	Fifth generation citations	710	629.54
	Total	1410	

Table 8: Independent-Samples Kruskal-Wallis Test Summary

Total N	1410
Test Statistic	68.58[a]
Degree Of Freedom	3
Asymptotic Sig.(2-sided test)	.00
a. The test statistic is adjusted for ties.	

Given that **Hypothesis 6_0** was rejected as there was a statistical difference between the semantic similarity scores between the generations of citation, pairwise comparisons between consecutive generations was examined using Bonferroni correction. The result of the pairwise comparison test is displayed in Table 9.

Table 9: Pairwise Comparisons of the Semantic Similarity Score categories

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Fifth generations semantic similarity score-Fourth generation semantic similarity score	101.69	25.46	4.00	.00	.00
Fourth generation semantic similarity score -Third generation semantic similarity score	86.55	35.26	2.46	.014	.085
Third generation semantic similarity score -Second generation semantic similarity score	99.53	49.87	2.00	.046	.276

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.

Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

[a] Significance values have been adjusted by the Bonferroni correction for multiple tests.

Cascading Citation and the Proposed Indirect Citation Weighting Comparison

The comparison between the cascading citation system and the proposed residual citation weights is in two phases. In the first phase, cascading citation weights were compared to the proposed indirect citation weights for each of the ten base articles at every generation. In the second phase, cascading citation weight per indirect citation was compared to the average semantic similarity score.

Cascading citation weights and the proposed indirect citation weights per base article

The cascading citation weights compared to the proposed indirect citation weights for each of the ten base articles at the second generation is visualized in Figure 8. A total of 20% of all the base articles received zero indirect semantic similarity-based citation weights, while all the base articles received equal cascading residual citations. Article 8 and article 9 received the highest residual semantic similarity-based citation weight of 5. Only two articles (article 8 and article 9) received the same value of cascading citation weights and semantic similarity-based citation weight. At least 80% of the residual citation weights of three base articles' (article 6, article 8 and article 9) second-generation articles were allocated non-zero semantic similarity-based citation weights. Nevertheless, the weights under the proposed method were lower than those of the cascading citation system, except on two occasions.

The comparison between the proposed citation weights and the cascading citation system at the third generation is visualized in Figure 9. At the third generation, the number of indirect citations to the base articles doubled, though the cascading citation weights remained the same. The number of base articles that got zero residual citations also increased in the third generation. Unexpectedly, the number of non-zero weights reduced from the second generation though the number of indirect citations increased at this generation as 50% of the base articles received zero weights when the lowest citation contexts' semantic similarity scores were considered for weight allocation. On the other hand, on two occasions,

base articles got more residual citations from the proposed method than the cascading citation when the highest citation contexts' similarity scores were considered for weight allocation.

The comparison between the proposed citation weights and the cascading citation system at the fourth generation is visualized in Figure 10. The number of indirect citations to the base articles quadrupled at the third generation, though the cascading citation weights remained the same. The number of base articles that got zero residual citations also increased in the third generation. The average value of residual citations per base article continued to increase, though the semantic similarity score average reduced.

The comparison between the proposed citation weights and the cascading citation system at the fifth generation is visualized in Figure 11. The number of indirect citations to the base articles increased eight folds at the fifth generation, though the cascading citation weights remained the same. The number of base articles that got zero residual citations also increased in the fifth generation.

Comparison between cascading citation weight and the highest semantic similarity score per second-generation article

Hypothesis 7_0, Hypothesis 8_0, Hypothesis 9_0, and **Hypothesis 10_0** were stated to guide the study. The hypotheses were stated to determine if the differences between ½, ¼, 1/8, and 1/16 (cascading citation weights) and the semantic similarity scores per second, third, fourth and fifth-generation articles, respectively. For instance, **Hypothesis 7_0** was stated to investigate if there was a significant difference between ½ (second-generation cascading citation weight) and the semantic similarity scores at the second generation.

Since the distributions of the semantic similarity weights are normal (see Figure 3, Figure 4, Figure 5, Figure 6), one sample T-Test was considered appropriate to investigate if there is are statistical differences between the semantic similarity scores and cascading citations. The result of one sample-T-test statistical test that was performed on the appropriate datasets to investigate the stated hypotheses is displayed in Table 10. The result showed that there is a significant difference between the cascading citation weight and the residual citation score at every generation. It was found there was a significant difference $t(99)=-2.47$, $p=.02$, between the cascading citation weight and average residual citation score at the second generation. It was found there was a significant difference $t(199)=20$, $p=.00$, between the cascading citation weight and average residual citation score at the third generation. It was found there was a significant difference $t(399)=46.13$, $p=.00$, between the cascading citation weight and average residual score at the fourth generation. It was found there was a significant difference $t(709)=75.41$, $p=.00$, between the cascading citation weight and average residual citation score at the fifth generation.

Hypothesis 7_0: There is no significant difference between the cascading citation weight and average residual citation score per second-generation article.

Hypothesis 8_0: There is no significant difference between the cascading citation weight and average residual citation score per third-generation article.

Hypothesis 9_0: There is no significant difference between the cascading citation weight and average residual citation score per fourth-generation article.

Hypothesis 10_0: There is no significant difference between the cascading citation weight and average residual citation score per fifth-generation article.

Table 10: One-Sample T-Test result for the Average Residual Citation Score

Sample 1-Sample 2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.[a]
Second generation residual citation-fourth generation residual citation	101.69	25.46	4.00	.000	.000
Fifth generation residual citation-third generation residual citation	188.24	32.60	5.78	.000	.000
Fifth generation residual citation-second generation residual citation	287.77	43.49	6.62	.000	.000
Fourth generation residual citation-third generation residual citation	86.55	35.26	2.46	.014	.085
Fourth generation residual citation-second generation residual citation	186.08	45.52	4.09	.000	.000
Third generation residual citation-second generation residual citation	99.53	49.87	2.00	.046	.276

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.

Asymptotic significances (2-sided tests) are displayed. The significance level is .050.

[a]Significance values have been adjusted by the Bonferroni correction for multiple tests.

Discussion

Research Question 1: How different is the proposed semantic similarity-based indirect citation weights from the cascading citation weights?

Similar to the cascading citation weighting, it was observed that the average semantic similarities values reduced as the number of citation generations increased. However, there were differences between these average values and those proposed by the cascading citation weights. The average for the highest semantic similarity score was lower than the average indirect citation from the cascading citation system at the second generation. In contrast, the average for the highest semantic similarity score was higher than the average indirect citation from the cascading citation system at the third, fourth and fifth

generations. This implies the change in the average semantic similarity from one generation to the next was not exponential in the proposed indirect citation weighting system, unlike the cascading citation system.

Since the proposed indirect citation is a derivative of the semantic similarities, it is only logical that the average residual citations accrued to the base articles reduced as the number of citation generations increased. It is interesting to note that the cascading citation system is limited in many ways. First, this study has revealed that some blanket values cannot determine the residual citation accrued to an article from its indirect citations. Residual citations are dependent on factors that could be behavioural; indirect citations are therefore dynamic, as shown in this study. From the result of this thesis, some articles were able to accrue indirect citation greater than the cascading citation weights; others accrued less.

Research Question 2: What differences exist in the highest semantic similarity scores between the generations of citation?

As expected, the similarity between the citation context of the first-generation citation and the second, third, fourth and fifth generations declined as the number of generations increased. This means, on average, the rate at which a publication potentially transfers knowledge to its indirect citations reduces as the generations increases. In corollary, the amount of residual citations due to an article reduced as the number of citation generations also increased. Like the cascading citation weighting system, the average amount of residual citation accrued from the indirect citation was highest at the second generation and reduced as the number of generation increased.

It was interesting to observe at individual base articles how this played out. The amount of knowledge transferred from the individual base articles reduced consistently from the second to the fifth in articles. Interestingly, for base articles that received an above-average residual citation at the second generation, the residual citation they received at the subsequent generations was also above average, and vice-versa. This means scientific publications possess different features, and possibly some factors determine the contributions of an article beyond its direct citations. While some consistently contribute indirectly above average and others do not, this is an area of research that deserves some attention in the future.

Research Question 3: What is the pattern of residual citations between the sampled cited documents and the nth generation citations?

How often does an article receive non-zero residual citation weights from its indirect articles? It was revealed that about 40% of the indirect articles produced non-zero weights at the second generation; this proportion reduced to 26.5%, 21%, and 10% at the third, fourth and fifth generations, respectively. This is not surprising given that it was initially noted that the semantic similarity scores reduced as the number of generations increased. The non-zero weights pattern throws more light into the amount of the indirect citations that received meaningful contributions from the base articles. The semantic similarities averages do not tell the whole story because they do not give the idea about how meaningful the contributions from base articles to indirect citations. For instance, this result shows the proposed

weighting system's dynamism instead of the linear system proposed by the cascading citation system. While the difference in the proportion of non-zero weights between the third and fourth generations seems close, further investigation is needed to ascertain the pattern of decline in the proportion of non-zero indirect citation weights as the number of generations increases.

Three categories of base articles were observed. Articles that received zero non-zero weights from all their indirect citations at all generations. The second category of base articles are averagely looking; they received average non-zero weights from their indirect citations across all generations. The third category of base articles received above-average non-zero weights from their indirect citations across all generations. The result showed that the base articles in the three categories that were described received either relatively high or low amounts of non-zero weights at every generation. This implies that beyond the behavioural factor that may impact the amount of residual citation that could be accrued to an article, some articles could be influential, probably because of the amount of information they contain. Therefore, very useful articles receive more residual citations than less useful articles. The proposed indirect citation weighting, therefore, is an important metric for weighting the influence of an article.

Conclusion And Recommendations

The proposed residual citation weighting method requires more complex computation than the cascading citation system. However, the proposed residual weighting system helps to fulfil the objective of residual citation allocation, which is to fairly quantify the attribution of scientific contributions to generations of citation on the citation path of a previously cited article. These so-called residual citations, i.e., the ones that are typically overlooked as a contribution by omission /attrition, in subsequent citations in the second, third or nth generations, are then reconstituted. Therefore, it is recommended that future studies compare computed results based on the proposed method to human judgement for the allocation of residual citations from scientific articles. Secondly, the proposed residual citation weighting is recommended over the cascading citation method because this method is based on the contribution of articles.

Declarations

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the Mendeley repository, through doi: 10.17632/6fgjxkv28d.3

Competing interests

The authors reported no potential competing interests.

Funding

This journal publication is part of the first author's doctoral thesis2 and contains texts that have been copied verbatim from the thesis.

Acknowledgements

The first author received the Western Graduate Research Scholarships from September 2016 and September 2020 and Ontario Graduate Scholarships and the Queen Elizabeth II Graduate Scholarships in Science and Technology (OGS/QEII-GSST), 2019 summer term to 2020 winter term.

References

1. Asubiaro, T. V. (2021). Exploiting Semantic Similarity Between Citation Contexts For Direct Citation Weighting And Residual Citation [Doctoral Thesis]. The University of Western Ontario.
2. Dervos, D. A., & Kalkanis, T. (2005). cc-IFF: A Cascading Citations Impact Factor Framework for the Automatic Ranking of Research Publications. 2005 IEEE Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 668–673. https://doi.org/10.1109/IDAACS.2005.283070
3. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., & Zhai, C. (2014). Content-based citation analysis: The next generation of citation analysis. Journal of the Association for Information Science and Technology, 65(9), 1820–1833.
4. Doslu, M., & Bingol, H. O. (2016). Context-sensitive article ranking with citation context analysis. Scientometrics, 108(2), 653–671. https://doi.org/10.1007/s11192-016-1982-6
5. Fragkiadaki, E., Evangelidis, G., Samaras, N., & Dervos, D. A. (2009). Cascading Citations Indexing Framework Algorithm Implementation and Testing. 2009 13th Panhellenic Conference on Informatics, 70–74. https://doi.org/10.1109/PCI.2009.30
6. Iqbal, S., Hassan, S.-U., Aljohani, N. R., Alelyani, S., Nawaz, R., & Bornmann, L. (2021). A decade of in-text citation analysis based on natural language processing and machine learning techniques: An overview of empirical studies. Scientometrics, 126(8), 6551–6599. https://doi.org/10.1007/s11192-021-04055-1
7. Jeong, Y. K., Song, M., & Ding, Y. (2014). Content-based author co-citation analysis. Journal of Informetrics, 8(1), 197–211. https://doi.org/10.1016/j.joi.2013.12.001
8. MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. Journal of the American Society for Information Science, 40(5), 342–349. https://doi.org/10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U
9. McCain, K. W. (2014). Assessing obliteration by incorporation in a full-text database: JSTOR, Economics, and the concept of “bounded rationality.” Scientometrics, 101(2), 1445–1459. https://doi.org/10.1007/s11192-014-1237-3
10. McKeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., Biran, O., Bothe, S., Collins, M., Fleischmann, K. R., Gravano, L., Jha, R., King, B., McInerney, K., Moon, T., Neelakantan, A.,

O'Seaghdha, D., Radev, D., Templeton, C., & Teufel, S. (2016). Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology, 67(11), 2684–2696. https://doi.org/10.1002/asi.23612

11. Merton, R. K. (1965). On the Shoulders of Giants_ A Shandean Postscript-Free Press (1965).pdf. The Free Press, New York.
12. Merton, R. K. (1988). The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. Isis, 79(4), 606–623. https://doi.org/10.1086/354848
13. Wallin, J. A. (2005). Bibliometric Methods: Pitfalls and Possibilities. Basic and Clinical Pharmacology and Toxicology, 97(5), 261–275. https://doi.org/10.1111/j.1742-7843.2005.pto_139.x
14. Wan, X., & Liu, F. (2014). Are all Literature Citations Equally Important? Automatic Citation Strength Estimation and Its Applications. Journal of the Association for Information Science and Technology, 65(9), 1929–1938. https://doi.org/10.1002/asi.23083

Figures

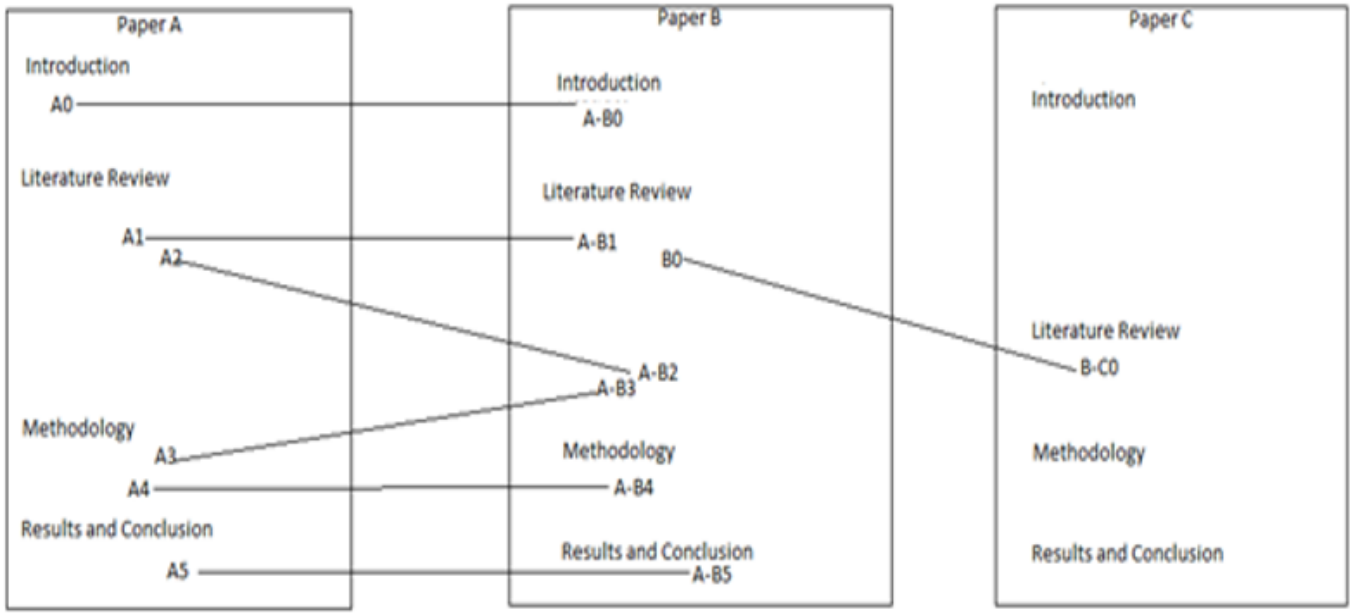

Figure 1

Three publications on a citation chain

Base1_FirstGen-article1_SecondGen-article1	in the united states approximately 63,000 new cases of t	the patient with a thyroid nodule should be asked about a fa	0.4799825	
	in the united states approximately 63,000 new cases of t	the recent 2015 american thyroid association management g	0.48888314	1
	in the united states approximately 63,000 new cases of t	recommendation 9 in the 2015 american thyroid association	0.47426349	
	in the united states approximately 63,000 new cases of t	the 2015 american thyroid association guidelines recommen	0.47357142	
	in the united states approximately 63,000 new cases of t	the appropriate level of thyrotrophine stimulating hormone s	0.45461038	
	in the united states approximately 63,000 new cases of t	therefore women with an excellent response to therapy as d	0.39346606	
	in the united states approximately 63,000 new cases of t	although fine-needle aspiration of subcentimeter nodules in	0.36851498	
Base1_FirstGen-article1-SecondGen-article2	in the united states approximately 63,000 new cases of t	the american thyroid association recently published updated	0.48337001	2
	in the united states approximately 63,000 new cases of t	there is a 75 response rate by 3 months and 89 rate by 1 yea	0.38524714	
	in the united states approximately 63,000 new cases of t	the nodule is rarely eradicated in patients with toxic adenom	0.37353265	
	in the united states approximately 63,000 new cases of t	thorough assessment of suspicious nodules within a toxic mu	0.37957019	
	in the united states approximately 63,000 new cases of t	both the american thyroid association and american associa	0.45154977	
Base1_FirstGen-article2-SecondGen-article3	lung breast prostate and colorectal cancer are considere	pancreatic cancer is the fourth leading cause of cancer death	0.61345756	
	the leading cancer sites in 2030 are predicted to be pros	pancreatic cancer is the fourth leading cause of cancer death	0.54136306	
	in 2010 and estimated for 2014 lung prostate and colore	pancreatic cancer is the fourth leading cause of cancer death	0.67011589	3
	thyroid cancer which is generally treated by surgical rese	pancreatic cancer is the fourth leading cause of cancer death	0.46988156	
	although there will be only an estimated 33,000 new cas	pancreatic cancer is the fourth leading cause of cancer death	0.53657889	
	pancreas cancer has the lowest 5-year relative survival r	pancreatic cancer is the fourth leading cause of cancer death	0.60530525	
Base1_FirstGen-article2-SecondGen-article4	lung breast prostate and colorectal cancer are considere	mortality due to pancreatic cancer is projected to surpass tha	0.66508615	
	the leading cancer sites in 2030 are predicted to be pros	mortality due to pancreatic cancer is projected to surpass tha	0.62943709	
	in 2010 and estimated for 2014 lung prostate and colore	mortality due to pancreatic cancer is projected to surpass tha	0.7278195	4
	thyroid cancer which is generally treated by surgical rese	mortality due to pancreatic cancer is projected to surpass tha	0.45002961	
	although there will be only an estimated 33,000 new cas	mortality due to pancreatic cancer is projected to surpass tha	0.61058056	
	pancreas cancer has the lowest 5-year relative survival r	mortality due to pancreatic cancer is projected to surpass tha	0.66014212	

Figure 2

Citation context pairs for four articles with marked highest semantic similarity measure

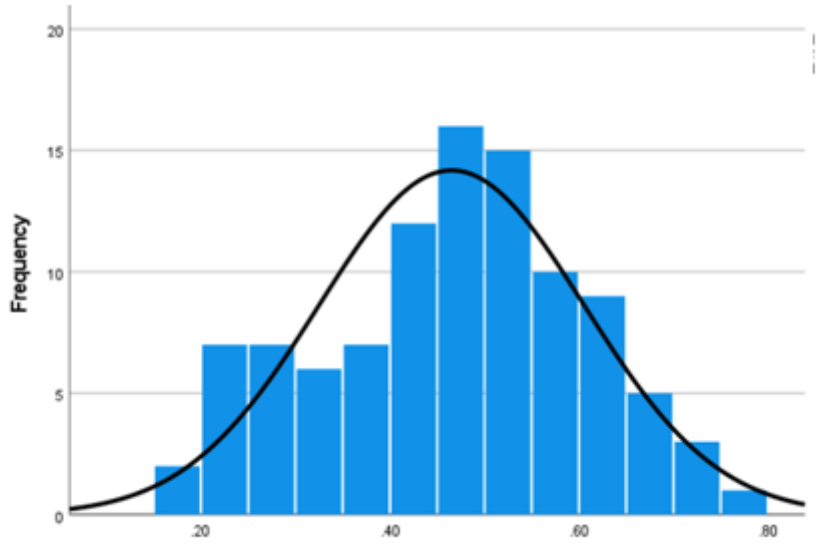

Figure 3

Distribution of the residual citations received from the second generation citations

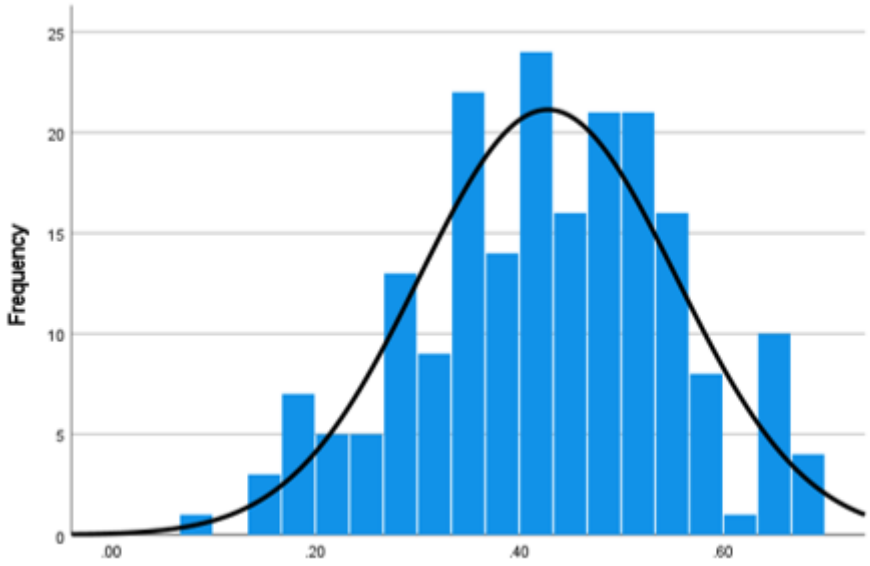

Figure 4

Distribution of the residual citations received from the third generation citations

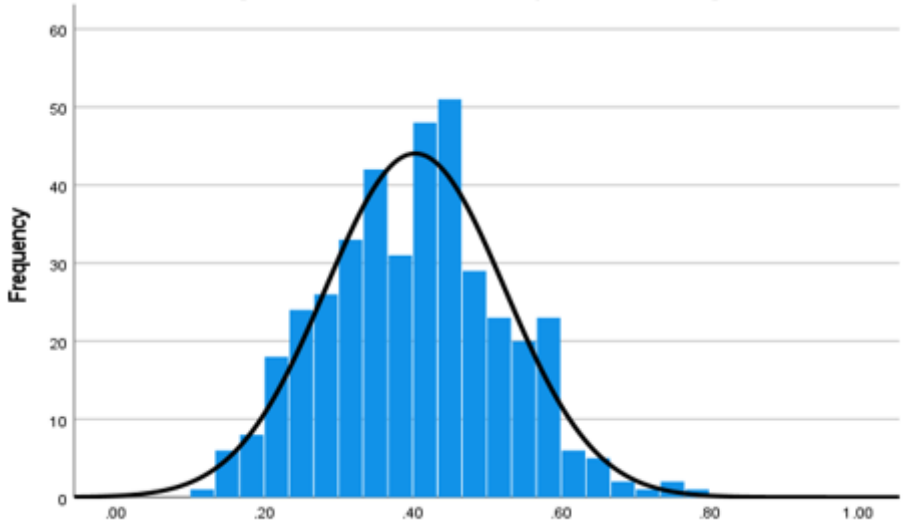

Figure 5

Distribution of the residual citations received from the fourth generation citations

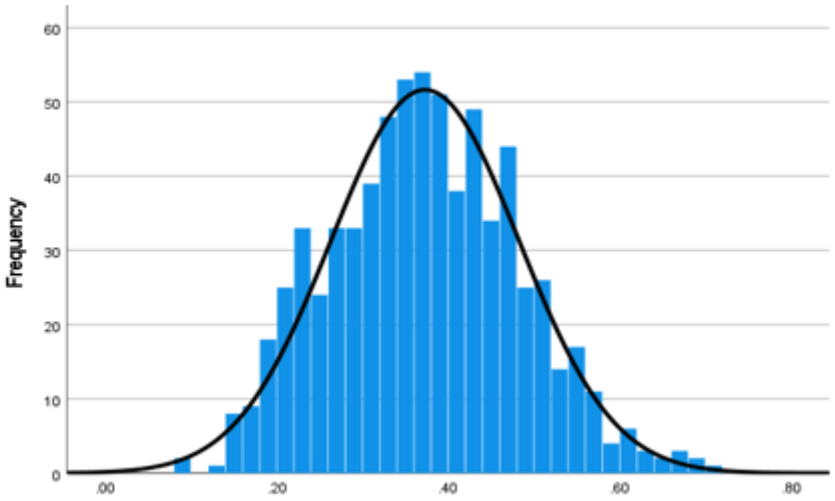

Figure 6

Distribution of the residual citations received from the fifth-generation citations Indirect Citation Weights

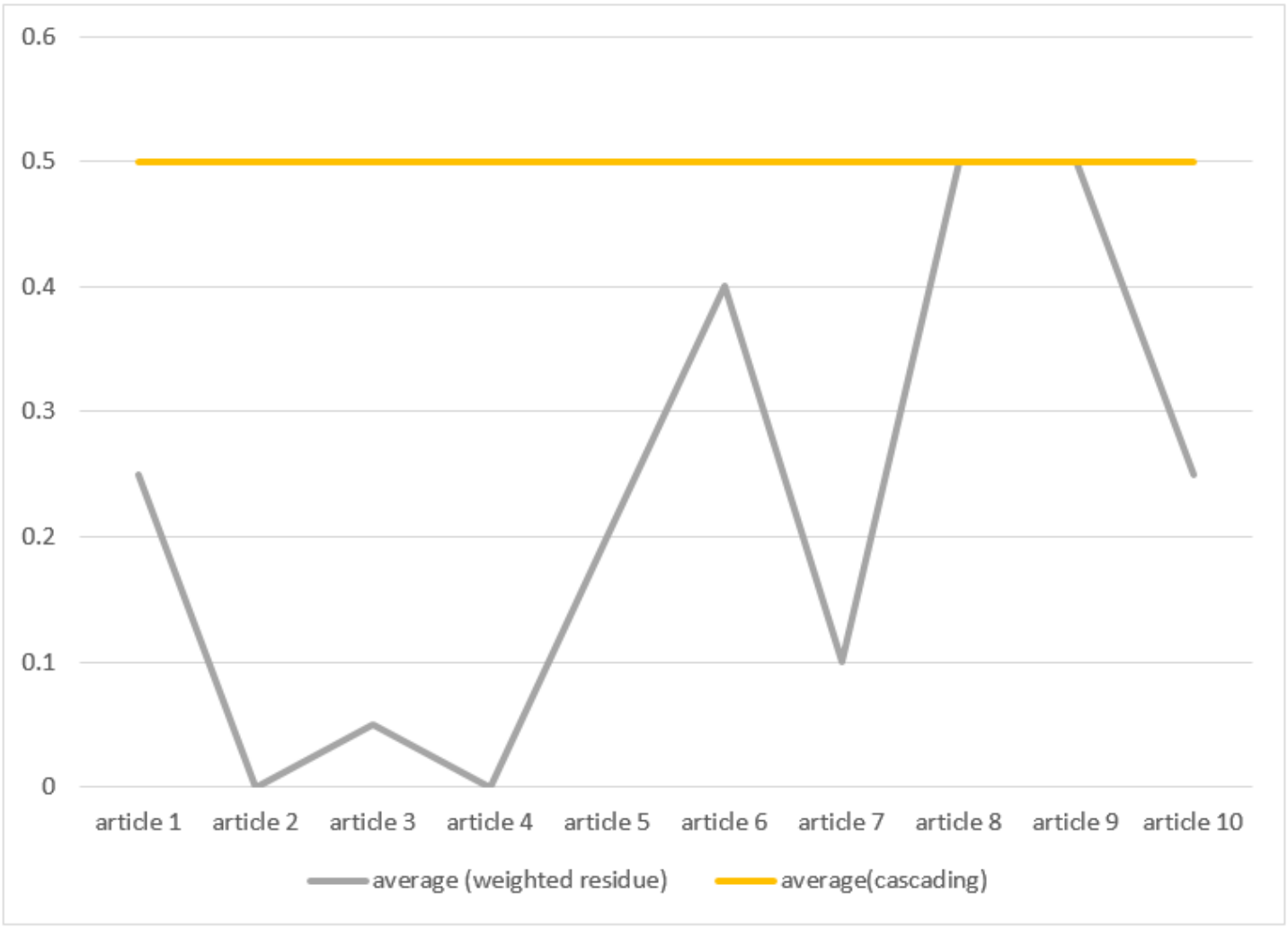

Figure 7

Second Generation Indirect Citation Weights Comparison

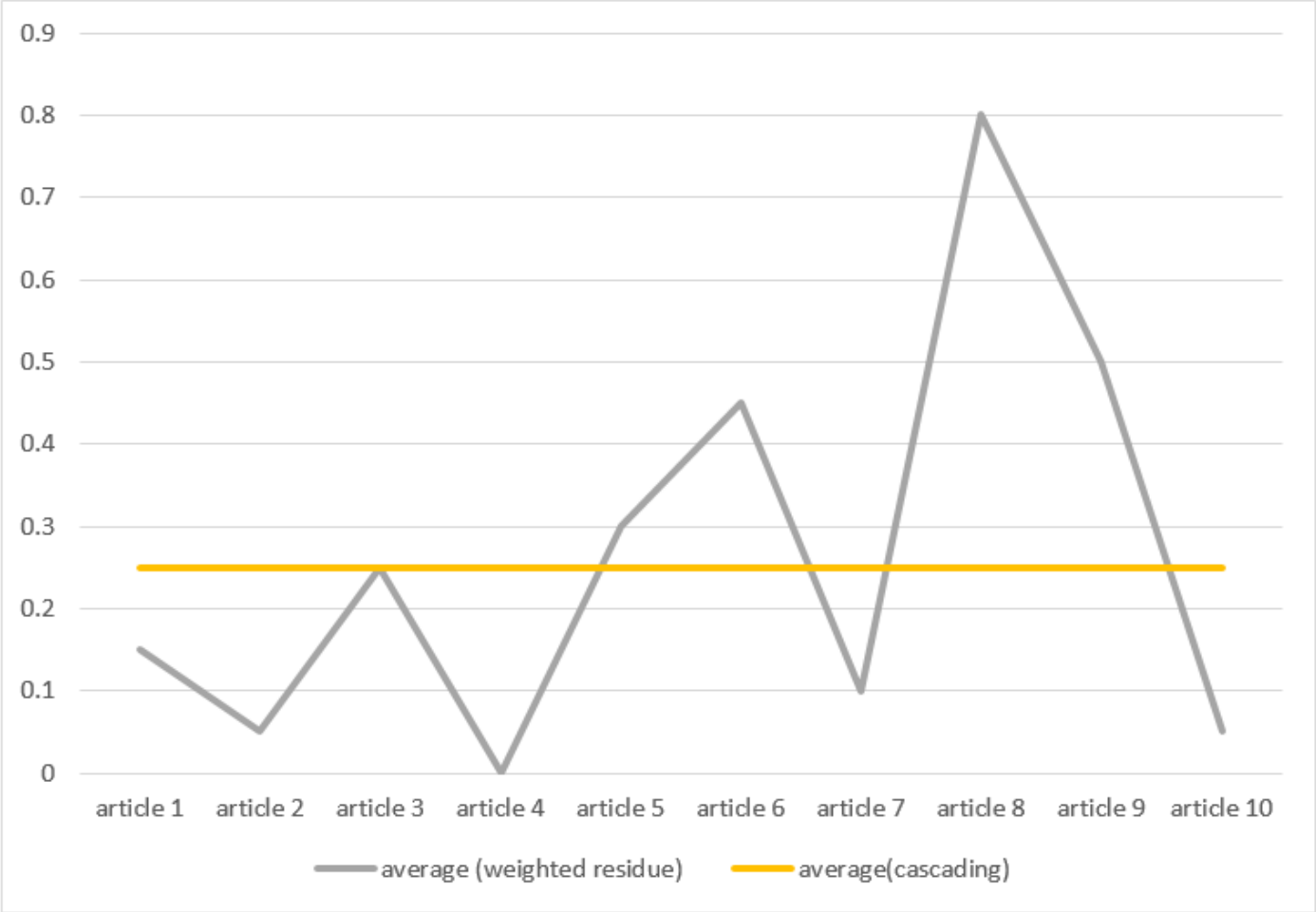

Figure 8

Third Generation Indirect Citation Weights Comparison

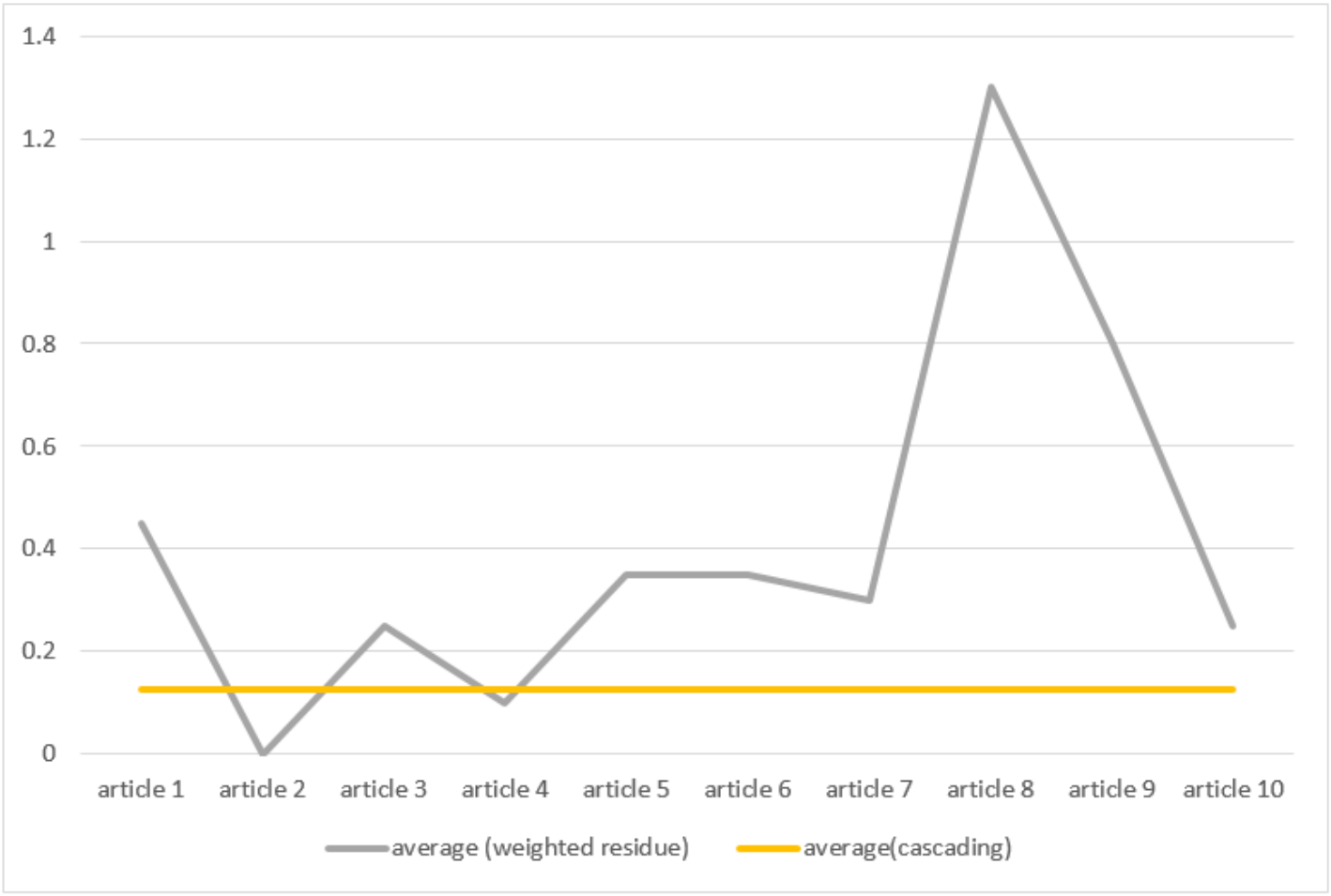

Figure 9

Fourth Generation Indirect Citation Weights Comparison

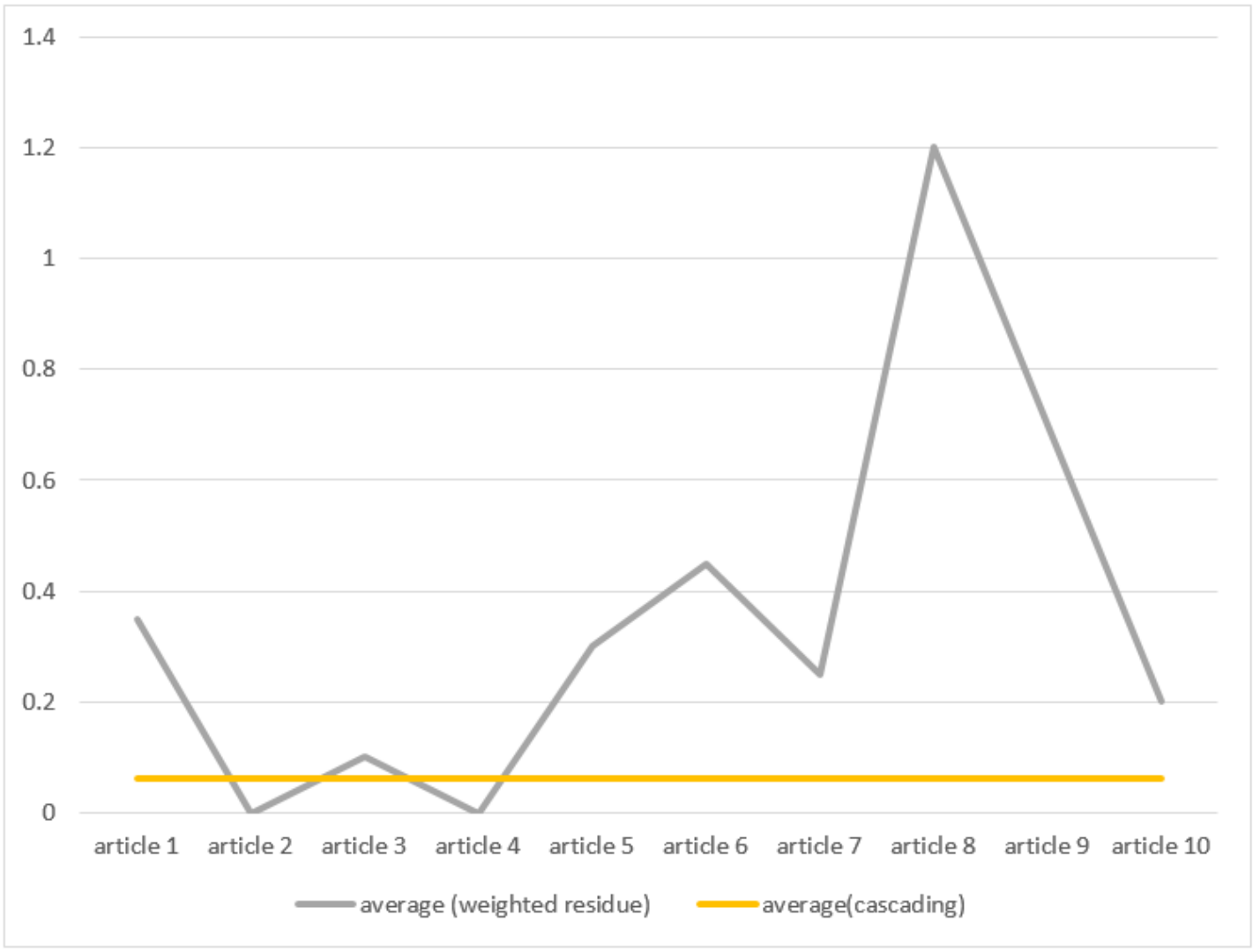

Figure 10

Fifth Generation Indirect Citation Weights Comparison