8-6-2021 1:00 PM

Exploiting Semantic Similarity Between Citation Contexts For Direct Citation Weighting And Residual Citation

Toluwase Victor Asubiaro, *The University of Western Ontario*

Supervisor: Ajiferuke, Isola, *The University of Western Ontario*

A thesis submitted in partial fulfillment of the requirements for the Doctor of Philosophy degree in Library & Information Science

Recommended Citation

Asubiaro, Toluwase Victor, "Exploiting Semantic Similarity Between Citation Contexts For Direct Citation Weighting And Residual Citation" (2021). *Electronic Thesis and Dissertation Repository*. 8008.
https://ir.lib.uwo.ca/etd/8008

Abstract

This study used the semantic similarity between citation contexts to develop one scheme for weighting direct citations, and another scheme for allocating residual citations to a publication from its nth citation generation level publication. A relationship between the new direct citation weighting scheme and each of five existing schemes was investigated while the new residual citation scheme was compared with the cascading citation scheme. Two datasets from biomedical publications were used for this study, one each for the direct and residual citation weighting aspects of the study. The sample for the direct citation aspect contained 100 publications that received 7317 citations, 11,234 citation contexts, and 9,795 citation context pairs. A sample of 981 citation context pairs was given to two human experts for annotation into "similar", "somewhat similar", and "not similar" classes. Semantic similarity scores between the 11,234 citation contexts were obtained using BioSent2Vec word-embedding model for biomedical publications. The residual citation aspect sample included ten base articles and five generations of citations from which 5272 citation context pairs were obtained. Results of the Spearman's rank correlation test showed that the correlation coefficients between the proposed direct citation weighting scheme and each of the weighting schemes "number of positive sentiments," "number of multiple citation mentions," "sum of multiple citation mentions," "number of citations," and "number of citation mentions" were .83, .89, .89, .93, and .99 respectively. The average residual citations received from the 2nd, 3rd, 4th and 5th citation generation level papers were 0.47, 0.43, 0.40, and 0.37 respectively. These average residual citations were significantly different from the averages of 0.5, 0.25, 0.125, and 0.0625 suggested by the cascading citation scheme. Even though the proposed direct citation weighting scheme and the residual citation scheme require more complex computations, it is recommended that they should be considered as credible alternatives to the "number of citation mentions" and cascading citation scheme respectively.

Keywords

Bibliometrics, Citation analysis, Citation Context Analysis, Citation Weighting, Indirect Citation Analysis, Residual Citation.

Summary for Lay Audience

One of the objectives of evaluative bibliometrics, a branch of Information Science, is to fairly and appropriately quantify the contributions from previously written (cited) papers to the citing scientific paper. Citation mention count, which is the number of times a cited publication is mentioned in the citing paper, is a popular method for weighting contribution of citations, it however does not take into account citation context information (the wording associated with the in-text citations). Firstly, this research proposes a more nuanced weighting method that incorporates citation contexts into citation mention count. Secondly, this study exploits the citation context information to create a system for weighting residual citation, where residual citations are accumulated by a publication depending on its contributions to other publications on its citation path. Conversely, on a citation path A-B-C, publication A was cited by publication B and publication C cited publication B, publication A contributed to publication C if the citation context of publication A in publication B is similar to the citation context of publication B in publication C. Two datasets were used for this thesis; one each for the direct and residual citation weighting aspects of the study. The first sample contained 100 publications that received 7317 citations, 11,234 citation contexts, and 9,795 citation context pairs. The proposed semantic similarity-based weighting allocated more weights to unique citation contexts. The indirect citation sample included ten base articles and five generations of citations from which 5272 citation context pairs were obtained. Statistical test revealed the number of citation mentions was the most similar metric to the proposed citation weight. This implies the proposed weighting method is similar to the citation mention method. Similar to the cascading citation system, residual citations received by articles from their generations of citations decreased as the number of generations increased. However, residual citations accrued to publications at all the generations were statistically different between the proposed and existing systems. This implies the proposed residual citation weighting is different from the cascading citation system. Though the proposed metrics require deeper computation, they are more novel because they are based on the contribution of the cited publications.

Dedication

To my late mother, an Amazon, Mrs. Felicia Aduke Bolatito Asubiaro (Nee Araoye) who sacrificed all to see me succeed. Abiyamo tooto.

Acknowledgments

I wish to acknowledge the help I received from God. The last Friday in August 2019, I walked into a clinic in London after a three-month wait for appointment to see a kidney specialist because I was diagnosed with a decline in my kidney function and high blood pressure six months earlier. The diagnosis was pronounced in February after a routine medical examination. The diagnosis was confirmed after a repeat of the kidney test, and three months later, the doctors said my kidney got worse. The statement in August, after six months, from the kidney specialist - “your kidneys are fine”, was a confirmation of a miracle without which I would not have finished my PhD. I am grateful to God’s for healing and keeping me alive.

I am grateful to my supervisor, Professor Isola Ajiferuke for his patience, kindness, guidance and contributions to this research work. He gave answers to many questions and provided leadership on many occasions. Ultimately, he allowed me to be creative. I want to specially acknowledge the assistance of Professor Victoria Rubin and Emeritus Professor Robert Mercer, members of my thesis committee and my teachers, who opened their doors for me at all times. I am also grateful to Emeritus Professor Catherine Johnson, Professor Kamran Sedig, Professor Jacquelyn Burkell, Grant Campbell and Professor Pam Mckenzie. My PhD journey could have been more difficult without these teachers turned helpers, mothers, fathers and friends.

I wish to also acknowledge my wife, she is the perfect, hardworking and dedicated (Phd) woman that helped made this happen. All my love to Akindamola and Toluwaniye, my

lovely sons, for their supports. Without your cooperation, love and sacrifice, it could have been impossible for us to achieve this feat. To my sister, Temitope Ayegbusi and her family, thank you. My unreserved gratitude to my Aunties, Olufunke Akomolafe and Odunayo Oni-Ogunleye and late Uncle, Emeritus Professor Akinyemi Araoye that passed away during my Phd program.

I am grateful for the support the following friends and families in Canada; Dele, Yetunde, Fikunmi, Kisi and Demilade Kilanko, Demola, Oyindamola and Dara Orekoya, Oluwole and Bimbola Badmus, Chukwuemeka, Isioma and Obi Okonkwo, Sarita Naa Akuye Addy, Marionette Ngole, Sarah Cornwell. I must drop a line to appreciate Dr. Benedict Oladele, the former University Librarian, University of Ibadan, Nigeria, who is always delighted in my progress. I am grateful for the support the following friends in Nigeria: Oluwaseun Egbodor and Elias Ejeh Elias,

I am also grateful to Pastor Kirby and Mrs Diane Hewlett and the Apostolic Faith Church, Kitchener, Pastor Al and Roach and the Revival Centre and Gathering Place, London.

Lastly, I will like to acknowledge the opportunity the government of Ontario and Canada gave me to study, live and work in this blessed and beautiful country. My gratitude goes to the leadership of the Faculty of Information and Media Studies for the support you provided during my PhD program. The following FIMS staff are specifically appreciated: Cindy Morrison, Sharon Waters, Brandi Borman, Lilianne Dang, Shelley Long, Sara Irons, Matt Ward, Charllotte McClellan and Manni Harrington (FIMS Graduate Library).

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

Included in the first section of this chapter is a background to the ideas of direct and residual citations in academia. A statement of the research problem is presented in the second section of this chapter. Other sections are the statement of research objectives, research questions and hypothesis in the third, fourth and fifth sections, respectively. The last three sections are the significance of the study, definition of terms and an outline of the thesis.

1.1 Citation in Practice

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). In other words, citations are considered accolades in scholarship, which can be accumulated by individual scholars, organizations, countries, or research outputs. Over time, these accolades can translate to economic, political and academic values. For instance, more cited articles are considered more valuable, and authors of such articles wield more political force in their fields of study. Besides, research works of the highly cited authors are more likely to be funded because more citations connote experience and trustworthiness in their areas of expertise. Also, criteria for the choice of academic prize winners are sometimes based on citation numbers as higher citation numbers are considered higher influence and impact. Lastly, highly cited journals are viewed to be of higher quality than journals that are less cited.

Quality assessment in the research community today is primarily built around citations and peer reviews. 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. Citations are academic commodities that are sold by researchers and bought by academic employers, prize panels and promotion boards for academic appointments, awards and promotions, respectively. Financially, citation analysis has an enormous stake as it is an “integral part of research quality evaluation and has been changing the practice of research” (Bornmann & Leydesdorff, 2014, p. 1228); more than one trillion dollars was spent on research and development globally in 2017 (Organization for Economic Co-operation and Development, 2018).

Citation analysis provides the most widely used methods for research evaluation (Bornmann & Leydesdorff, 2014). It helps to understand the “conceptual and professional evolution" of science (Larivière et al., 2012, p. 1000) through the study of diffusion of ideas (Sun et al., 2016; Zhao & Logan, 2002; Zhao & Strotmann, 2008). It also helps to understand the scope and characteristics of the scientific network and communities (Sun et al., 2016; Wagner & Leydesdorff, 2005). Other aspects include co-citation analysis, which is the study of citations that occur together. This is mainly used in studying the relationships between documents, research areas and researchers by analyzing the co-occurrence of citations in scientific communications. Co-citation analysis is also applied to document information retrieval and studying the features and evolution of scientific communities (Cottrill, Rogers, & Mills, 1989; Jeong, 2016; Jeong, Song, & Ding, 2014; Kim, Jeong, & Song, 2016). Bibliographic coupling, another aspect of citation analysis, is the study of the references that scientific papers share. This is applied to document ranking in information retrieval systems and also for quantifying the contribution of cited papers in the citing papers (Biscaro & Giupponi, 2014).

Famous metrics, which are based on the simple citation count, include the journal impact factor (JIF) which was proposed by Garfield (1972). The 2-year JIF of a journal in a given year, for instance, is the ratio of citations received by the journal in the year to the number of the citable items that were published in the journal in the preceding two years. For example, JIF for the year 2018 is the ratio of citations received by a journal in 2018 to the number of citable items that were published in the journal in years 2016 and 2017.

The EigenFactor (EF) score of the Thomson Reuters' Journal Citation Report (JCR) is a measure of the influence of a journal in relation to other journals in a field. EF is calculated as the citation number a journal has received from other journals (without counting self-citations) within a five-year window prior to a given year, divided by the number of citable articles that were published in the journal. The EF is identitical to the five-year JIF, except that citations are weighted while calculating the EF, where highly cited journals are weighted more. The Article influence score (AIS) is the normalized value of the EF by the total number of articles that have been published in a journal for five years (Bergstrom et al., 2008). H-index (Hirsch, 2005) and its variants like h2-index (Kosmulski, 2006), w-index (Wu, 2008), R and AR-indices (Jin, Liang, Rousseau, and Egghe, 2007) and g-index (Egghe, 2006) are metrics for ranking authors based on the citation numbers.

P-rank (Yan and Ding, 2010) is a citation count-based metric for measuring the prestige of articles, journals or authors where the citations received by an author, journal or article is weighted based on the number of citations received by the citation source. Therefore, citations from more cited journals, authors, or articles are allocated higher weights than those from less cited journals, authors or articles. The citation half-life is a metric for measuring the rate of obsolescence of journal articles (Burton and Kebler, 1960). The PageRank metric of the Google search engine, which is a weighted citation count-based metric, is used for ranking resources on the web, where more cited or more in-linked web pages or web articles are more weighted than less cited or less in-linked articles. Other popular search engines such as Microsoft Academic Search and CiteSeerX have also adopted citation count-based ranking systems for retrieval of scholarly articles (Wan and Liu, 2014).

Citation analysis, which is central to bibliometrics research, has many advantages over other research evaluation methods like peer review. First, datasets for bibliometrics research are easily obtainable where data is captured for a vast array of disciplines, huge samples (possibly almost exhaustible) and over long periods (possibly since the inception of science). This makes the bibliometric research more objective, less biased and reproducible (Asubiaro, 2018; Bornmann & Leydesdorff, 2014; Katz & Martin, 1997). Second, studies have shown that bibliometric indicators correlate well with other research evaluation metrics such as the number of external funding and scientific prizes won by researchers (Bornmann & Leydesdorff, 2014) and journals' ranking by experts' opinion (Sellers et al., 2004).

Thirdly, bibliometrics is concerned with the crème-de-la-creme of scientific work; publications in journals, conference proceedings and books are assumed to be research outputs that matter in the scientific community and unpublished results are deemed unimportant (Bornmann & Leydesdorff, 2014). Fourth, citation data is considered less biased because they are scientific artefacts which are produced by authors and are organized in bibliographic databases by independent bodies or computer algorithms. Authors, who are the citation data producers, are mostly not part of the bibliometric research processes, unlike producers of data for other research methods like surveys. Furthermore, the process of producing citation is embedded in a system that is largely reputable; citers "express their recognition and the influence of others' work" (Bornmann & Leydesdorff, 2014, p. 1228). Lastly, bibliometrics is a developed field with many standardized metrics, computer software, theories and methods (Asubiaro, 2018; Katz & Martin, 1997). Over the years, bibliometrics research has also applied methods and tools from fields like NLP, machine learning, content analysis and data science in a bid to produce more robust metrics.

Critics of the use of citation for research evaluation have identified some downsides. The downsides identified are in twofold: unethical authorship and fraudulent accumulation of citations (Bennett & Taylor, 2003; Teixeira da Silva & Dobránszki, 2016). Unethical authorship attribution includes gift, honorary, unjustified and guest authorship, which refers to the “inclusion of an individual in the by-line who does not meet authorship criteria” (Bennett & Taylor, 2003, p. 266). These can be in the form of the inclusion of the name of influential or senior academics as co-authors, with or without their knowledge or request. On the other hand, ghost authorship refers to “the failure to name an individual as an author when they have contributed substantially to the research or writing of the article” (Bennett & Taylor, 2003, p. 266).

Fraudulent accumulation of citation is achieved by practices such as “back-scratching” citations that occur among scholars that form “citation clubs”, a cycle of academics that cite themselves (Corbyn, 2008). Coercive citation, which is another citation malpractice, refers to subtle forceful means of making authors cite articles that may not necessarily improve their paper. Coercive citations are requested by some journal publishers or editors that ask authors to cite papers that are published in their journals as a pre-condition for manuscript acceptance (Fong & Wilhite, 2017; Wilhite & Fong, 2012).

Similarly, some crucial forms of research participation and engagements cannot be published, acknowledged in scientific publications, or rewarded with authorship. These engagements and participations include “casual interactions between researchers during which breakthrough ideas about a research work are mentioned”(Asubiaro, 2018, p. 29), or research ideas generated from discussions in a classroom classes (Glänzel & Schubert, 2004; Katz & Martin, 1997).

Furthermore, citation numbers are largely dependent on certain social, political, and technological factors. First, language plays an essential role in who gets cited by a potential research audience. Most studies in English are more cited because the audience for research written in English is larger than the audience for the studies in other languages. Second, some journals are not produced electronically and are not available to users on the internet; regardless of the quality or importance of the articles published in such journals, they are only likely to be cited locally. Third, the technological influence of document ranking affects citation. For instance, articles with accurate descriptions (metadata) are likely to be ranked higher, though the content may not be accurate or of quality. If such articles appear more in the first pages of search results, they are likely going to be read, utilized and cited more. Another technological influence is the ranking algorithm of popular search engines such as Google, which is influenced by the number of in-links to web pages (of scholarly articles alike), and it is designed to rank highly cited papers that are authored by more influential authors higher, thereby creating a loop. Less cited documents, authored by less influential authors, of more relevance and importance to a query may not be ranked high by the search engine.

Also, the simple citation count-based evaluation method assumes that all citations are equal. This does not reflect 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 the 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).

Citation context analysis incorporates the examination of citation contexts into citation analysis, and has been found useful in addressing some shortcomings of the traditional citation analysis and has profoundly influenced citation analysis field (Ding et al., 2014; Jeong et al., 2014). There are other citation weighting methods that do not require the inclusion of citation contexts, and these methods are mostly based on citation counts and theories from computer science and mathematics. For instance, the PageRank (Page et al., 1999) algorithm in Computer Science, which was proposed for ranking web pages, has been applied to citation-count based weighting (Fiala, 2012; Fiala et al., 2015; Fiala & Tutoky, 2017).

1.1.1 Citation Weighting Paradigm

Citation weighting, a paradigm shift from the traditional simple citation counting method, is based on the principle that citations are not of equal importance. Citation is either weighted for its contribution or impact in the content of the citing document relative to other citations or its antecedents, such as the influence of the journal from which the cited paper was published, the influence (citation numbers) of its author or its metadata such as the publication's age when cited. In the first instance, citation weighting refers to allocating a numerical value to the contribution(s) of a cited scientific paper in relation to other citations in the citing scientific paper in a fair and representative manner. Citation weighting based on the contribution of cited scientific papers is usually a complex and multi-dimensional problem. Citation weighting systems, therefore, have considered the inclusion of multiple dimensions of analysis to achieve more robust and objective systems. For example, citation weighting based on the dimension of function alone (without considering the citation sentiment, for instance), may not capture the effect of the negative, positive or neutral sentiments of the citations. Citation function, sentiment, frequency and location are some of the popular dimensions of analysis for weighting the contribution of citations. The inclusion of these dimensions or methods, and more, for citation weighting, will potentially present a more robust system.

Citation weighting methods that do not harness citation contexts are sometimes based on the source of citations' antecedents: that is, weighting is obtained by quantifying the relative fame or influence of the source (journal) or author of the cited scientific paper. It is assumed that a citation's contribution is directly proportional to the importance of its source (journal in which it was published) or author. Citation weighting in this category allocates more weight to articles from influential or famous journals or authors. Another method that is used for allocating weight is by considering the metadata of the cited article, such as its age at the time of citing, where it is assumed that more recent articles should be allocated more weights than older articles because current articles are perceived to be more important than older ones.

A more nuanced method for allocating weights to citations is through citation contexts or citation content analysis. Both citation context and citation content are used interchangeably in this thesis as they refer to the same concept. 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 also refers to the span of texts that represent the contribution of the cited publication in the citing publication (Doslu & Bingol, 2016). Citation context is usually part of or the entire sentence in which the in-text citation to the cited article in the citing article is located. The sentence in which the in-text citation is located is called the citation sentence. In addition to the citation sentence, citation context can also include a number of sentences before and/or after the citation sentence (Ding et al., 2014, p. 1821). Citation context identification, a non-trivial scientific assignment, is an endeavour to identify citation contexts.

Unlike the citation count, which allocates a count of one to the number of citations when a citing publication cites a paper, there could be more than one citation context of the cited publication in the citing publication. Each in-text citation of the cited publication is referred to as a citation mention. Therefore, the number of citation mentions of a cited publication represents the number of citation contexts of the cited publication.

Citation context analysis is concerned with the analysis of the citation context for allocating citation weights or classifying citations. Citation content analysis is a generic term, and it includes citation context analysis as it encompasses the analysis of some parts or the entire full text of the cited publication and/or citing publications for citation weighting or classification. Citation context analysis profer some solutions to a number of the shortcomings of ordinary citation count. For instance, while ordinary citation counts may not detect coercive citations, research has shown that citation weighting methods could be used for detecting coercive citations. Wilhite & Fong, (2012, p. 542) described coercive self-citation from journal editors to authors that submit manuscripts for peer-review as requests that (i) give no indication that a submitted manuscript is lacking in attribution; (ii) “make no suggestion as to specific articles, authors, or a body of work requiring review; and (iii) only guide authors to add citations from the editor’s journal”. Studies such as Yu, Yu, and Wang (2014), which focused on automatically detecting coercive citation, provide a shred of evidence that with citation context analysis, coercive citations could be detected automatically.

According to the literature, one of the most important methods in citation weighting is the citation mention analysis which allocates weights to citations based on the number of times a cited document is mentioned in a citing document (Zhu, Turney, Lemire, and Vellino, 2015; Boyack, van Eck, Colavizza, and Waltman, 2018; Sánchez-Gil, Gorraiz, and Melero-Fuentes, 2018). Studies have shown that the number of times a cited publication is mentioned is related to its contribution or importance in the citing document; about 75% of citations are mentioned once and perfunctorily (Stremersch et al., 2015; Zhao & Strotmann, 2014a). While the literature has shown citation frequency as a crucial syntactic feature for citation weighting, it is based on ordinary in-text citation count and ignores the citation context, which could be analyzed for more nuanced weights.

1.1.2 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 the scientific publications are benefiting 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, all publications are networked. On research networks, two publications are either connected by citation, co-citation or bibliographic coupling, the three defining factors for relatedness. Relatedness by citation refers to the relationship between two publications that is established by the citation, which suggest a level of relatedness between the citing and cited documents. Relatedness by co-citation refers to two documents that are cited together in the same documents and are therefore likely related. Lastly, bibliographic coupling occurs when two publications share a proportion of references, this is also a determinant of their level of relationship. The residual citation concept introduces a type of relatedness that is beyond these three traditional metrics, which is formed on citation paths. Citation paths refers 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 is a citation chain. The three existing forms of relatedness on citation networks are only defined between A-B and B-C, with no account for exploring relationship between A and C beyond the purview of bibliographic coupling and co-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. With the citation context analysis using computational methods, it is possible to quantify the level of relatedness between publications A-C based on the contribution of A in C, which could be antithetical to the cascading citation idea that allocates an equal value of citation residual to every indirect citation. While there is a robust literature on relatedness that is based on citation, co-citation and bibliographic coupling, none has explored the possibility of establishing the pattern of relatedness between articles on citation path that may not have been directly connected by citation.

1.2 Statement of the Research Problem

One of the objectives of citation weighting is to be fair in quantifying the contribution of a cited scientific paper in the citing scientific paper. While citation frequency analysis for citation weighting is simple, the fact that it is based on ordinary citation context count and texts of citation context is not considered in its computation suggests more nuanced weighting methods, that incorporate citation contexts analysis, could be created. Citation frequency citation weighting assumes that more frequently mentioned citations contribute more ideas and should receive more weights. There is a possibility that in some cases, the number of citation contexts may not accurately represent the contribution of the cited publications in the citing publication. Two cited articles may have the same number of citation mentions, but one may contain more unique contributions while the other only contains repeated contributions.

A citation context is a contribution of a cited article in the citing article. Therefore, a cited article that is referenced multiple times in a citing article is assumed to have contributed the number of times it is mentioned in the citing article, according to the citation mention weighting method. However, this thesis probes further by examining the possibility of allocating weights based on the uniqueness of contributions of the cited article. Thus, for the number of mentions of the cited article (marked by in-text citations), unique contributions are given more weights than repeated ones. There is a possibility that a citation that is mentioned multiple times in the cited article is only a repetition of citation contexts. Having a metric that is weighted based is on the uniqueness citation context is important because it gives a insight into the quality of contribution, as opposed to the quantity which is communicated through the citation mention analysis. This proposed method weights the contributions of the cited article in the citing articles better because the citation mention weighting method merely counted the number of contribution markers- the number of citation mentions, while our method places premium on the uniqueness of the contributions.

In practice, the idea of allocating residual citations to publications is fair. There are behavioural citation patterns that justify the allocation of residual citations to publications. For instance, some academics reference ideas that are part of in-text citations in scientific publication without giving credit to the original source; they only cite the paper in which they found the in-text citation context without citing the original paper. Similarly, the richness of a paper's contribution may be reflected in its nth generation citation. The state-of-the-art in allocating residual citation is based on cascading citation system that suggests all articles should receive equal residual citations from all their nth generation citations.

Though the cascading citation system provides a mechanism for accruing residual credits to scientific publications, this thesis argues that the cascading citation system is partly based on the conventional citation count idea, i.e., citations are equally weighed in each generation of citation. It is possible the cascading citation system is under-allocating or over-allocating residual citation to generations of citations based on its equal allocation of residual citations.

1.3 Research Objectives

The broad aim of this doctoral thesis is to propose a system for weighting citation based on the semantic similarity of the citation contexts and to explore the pattern of residual citations between sampled scientific publications and their five generations of citations. The specific objectives of the study are to:

1. create a system for weighting citations based on the semantic similarity of the citation contexts, and

2. investigate the knowledge flow pattern from a document to its second, third, fourth and-fifth generation citations.

1.4 Research Questions

The following research questions are intended to guide this study:

1. What is the relationship between the proposed semantic similarity-based citation context weight and existing metrics?
2. How different is the proposed semantic similarity-based residual citation weights from the cascading citation weights?
3. What differences exist in the residual citations among the generations of citation?
4. What is the residual citation pattern from cited documents and their nth generation citations?

1.5 Hypotheses

Hypothesis 1_0: There is no correlation between the number of citations and the proposed citation context similarity-based citation weight.

Hypothesis 2_0: There is no correlation between the number of citation mentions and the proposed citation context similarity-based citation weight.

Hypothesis 3_0: There is no correlation between the number of multiple citation mentions and the proposed citation context similarity-based citation weight

Hypothesis 4_0: There is no correlation between the sum of multiple citation mentions and the proposed citation context similarity-based citation weight

Hypothesis 5_0: There is no correlation between the number of positive sentiments and the proposed citation context similarity-based citation weight

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

Hypothesis 7_0: There is no significant difference between the cascading citation weight of ½ and the average residual citation score per second-generation article.

Hypothesis 8_0: There is no significant difference between the cascading citation weight of ¼ and the average residual citation score per third-generation article.

Hypothesis 9_0: There is no significant difference between the cascading citation weight of 1/8 and the average residual citation score per fourth-generation article.

Hypothesis 10_0: There is no significant difference between the cascading citation weight of 1/16 and the average residual citation score per fifth-generation article.

1.6 Significance of the Study

One of the objectives of evaluative bibliometrics in the Library and Information Science sub-field of bibliometics, is to fairly and appropriately quantify the attribution of scientific contributions and knowledge flow from previously written (cited) papers to the citing scientific paper. In more traditional bibliometric methods, such as citation analysis, and in more recent citation weighting methods like citation mention analysis, metrics are computed without taking into account citation contribution. This research proposes metrics that incorporate citation context information into citation mention analysis. While there are many existing metrics for evaluating research, one of the potential benefits of this thesis is the proposed metrics which would potentially make attribution more fair because they are based on the contribution of the cited publication in citing publication. New metrics generate new discussions and studies; these metrics have the potential of being studied extensively for improvements, and complementing existing ones in evaluating research.

Secondly, much has been written in literature on contribution of citations to the citing documents. However, there is no information on the contributions of publications to indirect citations on their citation paths, though the evidence of certain citation habits suggest that studying contributions to indirect citation is an important aspect of evaluative bibliometrics. The results from this thesis will provide the information about the possibilities of contribution to generations of citations and how this type of contribution changes over time. This will potentially open a new discussion in bibliometrics, Information Science field and in academia about the phenomenon of residual citation allocation through the weighting of contribution of a publication.

Thirdly, citation context analysis is a budding research area with developments on the automatic identification of citation contexts. State-of-the art large scale studies use citation contexts that comprise a pre-specified window of text around citation markers, while studies have shown that a specified window of texts does not accurately represent citation because it can either over-represent or under-represents the citation contexts. These studies rely on this method because there are no clear-cut computational models or algorithms for accurately identifying citation contexts. One of the potential benefits of this study is the datasets which will be made available publicly for research. These datasets can be used for computational modelling, and other citation context studies.

1.7 Definition of Terms

1. Unweighted Citation: Citation counts in which every citation in the reference list is allocated the weight of one regardless of the contribution of the citation to the citing paper or the importance of the antecedent of its source.
2. Weighted Citation: Citation count system in which citations are not allocated the equal weights. Citations are either allocated weights based on their contribution or the importance of antecedents of its source.
3. Citation graph: A citation graph is “a representation of the relationships that exist between research articles based on the references that each article provides.” On a citation graph, articles are the nodes (Fragkiadaki, Evangelidis, Samaras, and Dervos, 2010)
4. Citation path: Citation path refers to the linear relationship between two articles that are related by citation, but not directly connected by citation. For instance, in Figure 1, A-B-I is a citation path, A-C-E-H is another path. Citation paths are normally acyclic.

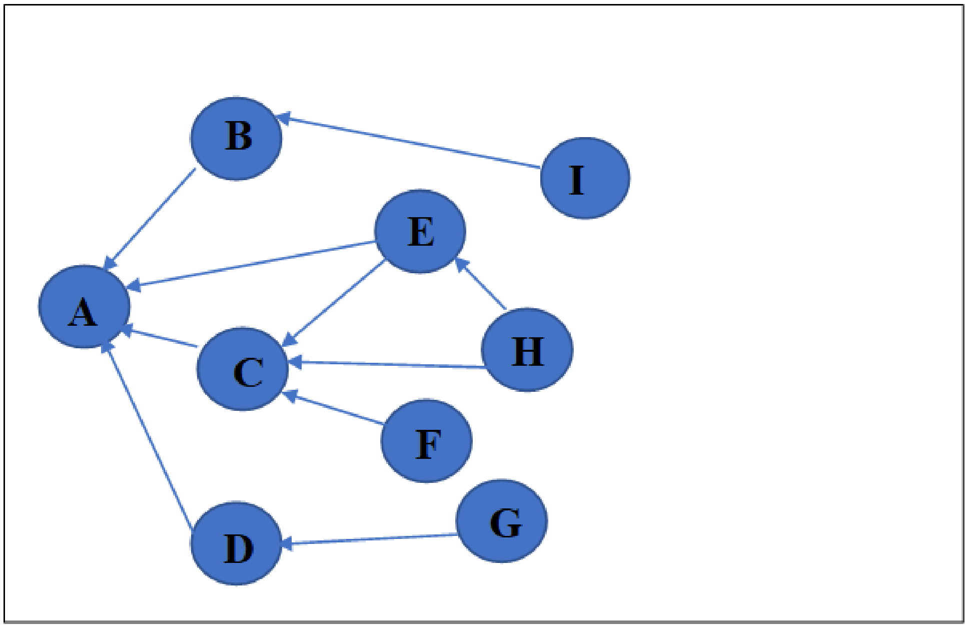


Figure 1.1: Citation network sample

5. Nth-generation citation: Nth Generation citation refers to the least number of nodes between a cited article and a given article that exist on its citation path. For instance, between A and B, there is only one. Therefore A receives 1st generation citation from B. First-generation citations are referred to as direct citations while subsequent generations are indirect citations.
6. Residual Citation: This is the credit attribution received by a publication from its second-generation citations to its nth generation citations.
7. Knowledge Flow: Knowledge flow is assumed to have occurred from a publication to its nth generation citation when the citation context of its n-1th generation in the nth generation citation is semantically similar to the citation context of the publication in its direct citation.

1.8 Outline of the Thesis

This thesis consists of six chapters. The first chapter provides a background to the subject of citation, its flaws and importance to scholarship. An introduction to the citation weighting paradigm is also discussed in detail, preparing readers for the depth of the report. Chapter one also contains a statement of research problems, an explanation of the research gap this thesis work fills. The research objectives, research questions and hypotheses are also presented in Chapter One. And lastly the significance of this thesis is presented in the first chapter.

Chapter two is the literature review chapter. Literature on citation context analysis, citation context classification and citation context weighting was reviewed in the first three sub-sections of chapter two. Literature on the existing citation weighting frameworks was explored in the fourth sub-section. Proposed frameworks for direct and indirect citation weighting were also presented in the second chapter. Lastly, the contribution of this study is presented.

The third chapter presents the methodology of this study. The first sub-section in the methodology chapter presents the data sampling method for the direct and indirect citation analysis parts of this study. The second sub-section of the third chapter contains text extracting data-citation contexts- from the full texts of the sampled scientific publications. The data pre-processing sub-section includes the text cleaning and preprocessing steps that were taken. Steps taken to annotate a sample of the direct citation weighting dataset by human experts is presented in the fourth sub-section of the methodology chapter. An algorithm for obtaining the semantic similarity between citation contexts, based on the cosine value between their vectors, was presented in the third chapter's fifth sub-section. The three implementations of the proposed methods for weighting direct citation contexts were presented in the methodology chapter's sixth sub-section. The strategy for indirect citation context weighting and analysis was presented in the methodology chapter's last sub-section.

The results of the analysis are presented in the fourth chapter. The results are presented in four major sections of the fourth chapter. The result of the human annotation of a portion of the direct citation context dataset is presented in the first sub-section of chapter four. The second sub-section contains the description of the direct citation datasets. The analyses of the direct and indirect citation contexts are presented in the third and fourth sub-sections of the fourth chapter, respectively.

Chapter five contains the discussion of the salient results from the analysis. The results were discussed under the four research questions guiding this study. Therefore, the discussion section has one sub-section for the direct citation weighting aspect of this thesis. The last three sub-sections are discussions on the results of the residual citations aspect this research work.

Chapter six, the last chapter, contains the conclusion and recommendation. The sixth chapter is divided into four major parts. The first part of the sixth chapter contains a summary of findings from the study. The second part contains the conclusion to the doctoral work. The third part contains the recommendations for improvement and suggestions for further studies on the subject of direct and indirect citation weighting. The last sub-section of chapter six presents the limitations of this thesis.

Chapter 2

2 Literature Review

The literature review chapter is divided into eight sections. The first section reviews citation classification literature. The second section reviews the literature on citations weighing schemes and frameworks. The third and fourth sections focus on citation weighting and citation weighting frameworks, respectively. The fifth section presents an overview of citation context in citation weighting, while the cascading citation system is reviewed in the sixth section. The last two sections present the proposed frameworks for direct and residual citation weighting.

2.1 Citation Context Analysis

Citation context is the span of texts that surround the citation marker; used for referring to cited publications, 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. To identify the citation context, studies have suggesting various windows of text to capture the extent of the contribution of the cited publication in the citing publications. In the review of the citation context window size by Iqbal et al., (2021), it was revealed that range of one to four sentence-window and

50 to 100 word-window have been 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, while 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. While citation polarity is a continuous variable with values between +1 and -1, citation sentiment is the classification of citation polarity 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.

Syntactic elements that describe a citation are consolidated in the structure of scientific communication that mostly follows specific patterns. For instance, scientific communications are mostly presented in the introductory, methodology, discussion, recommendation and conclusion sections. Other sections are the article title, abstract, keywords, and bibliography. Considering these syntactic elements in processing citation data has provided new insights into citation classification research. Theoretical studies have propounded that the structure of scientific communications is useful in classifying citation contexts. Therefore, some aspects of citation context analysis are dedicated to analyzing citation context location for allocating weights and categories.

Studies in the 1970s and 1980s, such as Herlach (1976), Peritz (1983), and Voos and Dagaev, (1976) considered citation location and frequency in citation classification.

Recently, other features apart from the citation location and frequency, such as citation context lengths, citation intent, author overlap, number of direct citations, number of indirect citations, and PageRank of the cited article (Pride & Knoth, 2017; Valenzuela et al., 2015a), have been considered for citation classification in the literature. While the older studies were either done manually (Voos & Dagaev, 1976) or based on supervised machine learning methods (Teufel et al., 2006; Zhu et al., 2015), later ones such as Nazir et al., (2020), and Wang et al., (2020) used unsupervised machine learning methods such as neural networks, multiple regression analysis, support vector machine, random forest, and KNN for citation context classification.

Later works have studied syntactic features embedded in the linguistic structures of scientific communication. The theoretical background for the studies by Di Marco et al. (2006) and Mercer et al. (2004) was based on the idea that hedging (expressions that make statements more fuzzy) cues that are most familiar in citation contexts could be used for citation purpose classification. Using a catalogue of regular expressions of hedging cues, categories of citation contexts were identified, confirming that stylistic and rhetorical structure in scientific communication is useful in citation context classification. Citation function classification (Cohan et al., 2019; Teufel et al., 2006), citation intent classification (Dong & Schafer, 2011), citation importance classification (Qayyum & Afzal, 2019) were computed using other patterns such as subject, quantity, frequency, tense, example, suggest, hedge, idea, basis, comparison and result cues.

Metadata have also been used to investigate the relationship between citing and cited documents. For instance, Bonzi (1982) investigated the features of scientific publications that can determine relatedness between the cited and citing papers in LIS. Metadata such as article type, citation source, date of publication, sex of author, article type, article length, number of citations, many mentions of articles, placement of citations in articles, number of citation footnotes and journal type were features that could determine the relationship between the scientific articles. Only the journal type, article type and multiple mentions of citations were statistically significant in the analysis.

Semantic citation context analysis is usually carried out between some portions of the full texts of the cited and citing articles on the premise that the degree of similarity between the cited and citing document determines the cited article's contribution to the citing article. Semantic similarity between portions of the citing and cited articles’ full text such as title, abstract, complete full text, introduction sections, and conclusion sections in citation context analysis studies are essential. Zhu et al. (2015) used the semantic similarity between the title of a cited paper and the title, introduction, conclusion, and abstract of the citing article in determining the contribution of the cited article in the citing article. Pride and Knoth (2017) and Hassan et al. (2017) concluded in their studies that the best feature for citation classification is the similarity between the abstract of the citing and the cited paper. Semantometrics, a citation influence metric, in Knoth and Herrmannova (2014) was based on the premise that a citation context’s influence is proportional to the semantic distance between the cited article and its citing article.

2.2 Citation Classification

Basic quantitative or bibliometric features such as citation numbers, age, number of authors, author rank and citation mention count remain the most popular input for citation classification studies. From these basic bibliometric features, there are natural citation classes such as self-citation and external citations. Similarly, the bibliometric features are combined with qualitative features from the citation contexts in more complex citation classification studies. Other bibliometric metrics, such as the average citation score, the JIF, AIS, EF h-index and PageRank, which are derived from citation numbers, are mostly used in the literature for creating weighted metrics like the Field-Weighted Citation Impact (FWCI) (Colledge, 2014), co-author weight coefficients (Zhang, 2009), weighted citation (Yan & Ding, 2010). While citation weights are mostly represented as continuous data, citation classes are ordinal or categorical, though citation classifications are sometimes represented as weights. Qualitative features of a citation for classification are generated from the content or context in which citation appears in the citing text and sometimes in relation to the cited paper.

2.2.1 Citation Importance Classification

Studies have classified citation contexts based on their impact, contribution, importance or influence in the citing article. The essence of this type of classification is not to classify based on function or sentiment; rather, the distinction in the classes are meant to reflect the contribution of the citations. However, some of the classes in citation function studies reflect differentials in citation context contributions because the objective of such studies is not to illuminate the contribution of the citations; not all the classifications reflect differences in impact.

Different citation function classification schemes were proposed in the literature. The binary citation functions are the most popular, and classification schemes are based on importance or influence. Maričić, Spaventi, Pavičić, & Pifat-Mrzljak, (1998) adopted a binary citation function scheme; that is, cursory and meaningful citation functions. The cursory and meaningful citations are named differently in other studies with more citation function classes. According to Maričić, Spaventi, Pavičić, & Pifat-Mrzljak (1998) the cursory citations were classified as non-essential in Cano (1989), perfunctory in Moravcsik and Murugesan (1975) and peripheral in McCain and Turner (1989). The meaningful citations were classified as essential in Cano (1989), organic in Moravcsik and Murugesan (1975) and central in McCain and Turner (1989). One of the important results of Maričić, Spaventi, Pavičić, and Pifat-Mrzljak (1998) is that citations in the Introduction section of scientific publications are perfunctory, while citations in the methodology, discussion, and results sections are important.

Other studies which have developed binary classification schemes for citation function are Hassan et al. (2018), Valenzuela et al. (2015) and Hassan et al. (2017). While Hassan et al. (2018), and Qayyum and Afzal (2019) classified citation functions as ‘important’ and ‘non-important’, Valenzuela et al. (2015), and Hassan et al. (2017) created the ‘important’ vs the ‘incidental’ classes, where the incidental classes included the ‘related work’ and ‘comparison’ sub-classes and the ‘important’ classes included the ‘using the work’ and ‘extending the work’ sub-classes.

One feature that has been used to determine the importance of a citation is the location of the reference in the citing paper. Citation location analysis in citation classification assumes that citations in some sections of scientific write-ups contribute more than citations in most sections of a scientific paper. Paper sections that contain the fewest citations are the results section, but these sections will likely contain the most important citations. Herlach (1976) and Voos and Dagaev (1976) are two of the earliest studies that considered citation location analysis in finding out the contributions of cited documents in citing documents. These studies, though manually done and using a relatively small sample size, submitted that the fewest citations were located in the results section.

Citation locations have been mapped to citation impact. Maričić, Spaventi, Pavičić, & Pifat-Mrzljak, (1998) worked on identifying cursory and essential citation in the different sections of scientific publications and reported that the introductory section contains more cursory citations while other sections like the results, discussion and methodology contain fewer but more important citations. In this case, studies such as Hassan, Akram, and Haddawy, (2017), Pride and Knoth, (2017), An, Kim, Kan, Chandrasekaran, and Song, (2017), and Ding et al., (2013) showed that citations in the introductory and results/discussion sections are perfunctory or unimportant while in-text citations that were in the abstract or methodology are more important. Some studies like An, Kim, Kan, Chandrasekaran, and Song, (2017), and Ding et al. (2013) have provided descriptive analyses of the distribution of in-text citations in the full text of articles based on their locations (paper sections), and these studies have shown that citations are unequally distributed in the body of the texts of articles.

2.2.2 Citation Function Classification

Citation function studies investigate the “whys” a paper was cited. In some articles in the literature, citation function is also referred to as citation motivation, intention (HernáNdez-Alvarez & Gomez, 2016) or utility (Stremersch et al., 2015). In citation function classification studies, citations are grouped by classes based on the motivation, reason, or use of citations. The task in citation function study is to unravel the reason or intention of the citing author (Teufel et al., 2006) or how the cited work was used by the citing author (Tuarob et al., 2020). Most of the citation function classification schemes do not provide classes based on the importance or the impact of the papers that are cited. Rather, the classes give an idea of the different motivations for citing; a behavioural analysis. According to Moravcsik & Murugesan, (1975, p. 88), the distinctions in the classes are “not meant to be a value judgement, and are not to be taken as synonymous with judging the importance of the paper referred to.” For instance, of the citation functions classes use/application, affirmation/support, review, negation, and perfunctory citation functions (Baumgartner & Pieters, 2003), it could be argued that perfunctory or negation citation function classes are less important than others. However, these classes were created to capture the use of the references in the articles that cited them. Negation citation function could arise from a citer’s challenge of the ideas in a cited article, which could result in a study based on the cited article. In this case, several mentions of the cited articles could occur in the citing article, mostly negating the cited article. The negated citation may be more influential than citations in other citation functions classes.

Citation function studies use surveys and content analysis of publications as the two primary data collection methods. Few citation function classification studies collected survey data from researchers and the focus was on citation motivation. Survey data collection method is limited because of small sample sizes. Sample sizes in content analysis of publications are much larger due to improved computational methods for data science. For instance, Maričić, Spaventi, Pavičić, and Pifat-Mrzljak (1998), one of the oldest studies on citation function used about 300 publications for their study, while McKeown et al., (2016), performed a large-scale experiment with 3.8 million full-texts.

Two classes of citation functions are the least in the literature. Meyers (2013) suggested two classes of citation functions: contrasting and corroborating citation functions. The two citation function classes were based on how the cited and citing documents compare on the ideas or approaches. While the contrasting citation function category indicates a situation where the citing article may describe approaches or opinions different from the cited article's, on the other hand, corroborating citation function indicates a situation where both citing and cited articles follow the same approach. Tuarob et al. (2020) proposed a binary citation function scheme for algorithm citations in computer science publication, with *utilize* and *nonutilize* categories. The categories were based on the use or non-use of an algorithm from the cited publications. *Utilize* citation category was further sub-divied into *use,* and *extend*, while the *nonutilized* category was sub-divided into *mention* and *not an algorithmic citation context.*

Cohan et al. (2019) classified citation intent into three classes: background citation, method citation and result extension citation. Some three-class citation functions have been mapped to polarity (positive, negative, neutral) (HernáNdez-Alvarez & Gomez, 2016). For example, Li et al. (2013)'s citation function classification scheme contained three categories (positive, neutral, and negative), each of the three categories were further divided into subcategories. Positive citation function was further divided into *based on, corroboration, discover, positive, practical, significant, standard,* and *supply*. Neutral citation function was further divided into *contrast, co-citation* and *neutral* citation function classes. Teufel et al. (2006) mapped their 12-category citation function classification into three sentiment polarities. Weak (weakness of cited approach) and co-co (author's work is stated to be superior to cited work) categories were mapped to negative sentiment polarity. PMot (this citation is positive, about approach used, or problem addressed in the cited paper), PUse (author uses tools/algorithms/data/definitions), PBas(author uses cited work as basis or starting point), PModi (author adapts or modifies tools/algorithms/data), PSim (Author's work and cited work are similar), PSup (author's work and cited work are compatible/provide support for each other), were mapped to positive sentiment polarity. CoCoGM (contrast/comparison in goals or methods(neutral)), CoCoR0 (contrast/comparison in results (neutral)), CoCoXY (contrast between 2 cited methods), Nuet (neutral description of cited work, or not enough textual evidence for prior categories, or unlisted citation function) were mapped to neutral sentiment polarity.

Meng et al. (2017) adopting Dong and Schafer (2011), proposed the following four categories of citation functions: background, fundamental idea, technical basis, comparison. Jurgens et al. (2016) created a citation function classification scheme of six categories- background, motivation, uses, extension, continuation, comparison or contrast and future.

Using the Information and Science and Technology journal (JASIST) articles, Tabatabaei, (2013) categorized citation functions by research impact into five; ‘applied,’ ‘contrastive,’ ‘supportive,’ ‘reviewed’ and ‘perfunctory.’ The classification by Tabatabaei (2013) has some locational annotations, though the citation functions were based on their impact in the citing publication. The ‘applied’ citation function included citations that were mentioned in the analysis approach, the research model or theoretical framework sections of the citing documents. The ‘applied’ citation function also includes citations regarding the data, concepts, software/algorithm, criteria specified, or the hypothesis stated in the cited paper. ‘Applied’ citation function also includes the citations that are made in “Continuation/expansion/modification of previous studies.” ‘Contrastive’ citation function can either be “comparative,” “affirmative” or “critical.” ‘Supportive’ citation functions are cited as part of the methodology, findings, assumptions, research purpose, data, sample size, algorithm and further research suggestions sections. Zhao, Strotmann and Cappello, (2018) adopted the citation functions that were identified by Tabatabaei (2013) for categorizing self-citations in JASIST articles and reclassified the ‘reviewed’ and ‘perfunctory’ citation functions as non-essentials.

Peritz (1983) developed a more complex classification scheme of eight categories for citation motivation for empirical studies in the social sciences. The eight citation motivations or functions developed were: ‘setting the stage,’ ‘background,’ ‘methodological,’ ‘comparative,’ ‘argumental’/’speculative’/’hypothetical,’ ‘documentary,’ ‘historical’ and ‘casual.’ Abu-Jbara, Ezra, & Radev, (2013) classified citation function or purpose into six categories. The categories include ‘criticizing’ which could have ‘positive’ or ‘negative’ polarity, ‘comparison’ which overlaps with the ‘comparative’ category in the Peritz (1983) classification scheme, and use which refers to citations that use methods of the cited work. Other classes are ‘substantiating’, which implies that the results or claims of the cited documents are referenced, ‘basis’ when the citing paper cites the cited document as a motivation or starting point and ‘neutral’ when the citing cannot be categorized under any of the prior classes.

The most granular citation function classification scheme was proposed by Garzone and Mercer, (2000). Thirty-four citation function types in ten categories were proposed (see Table 2.1).

Table 2.1: Citation Function Classification Categories in Garzone & Mercer, (2000)

	Negational Type Categories
1	Citing work totally disputes some aspect of cited work.
2	Citing work partially disputes some aspect of cited work.
3	Citing work is totally not supported by cited work.
4	Citing work is partially not supported by cited work.
5	Citing work disputes priority claims.
6	Citing work corrects cited work.
7	Citing work questions cited work.
	Affirmational Type Categories
8	Citing work totally confirms cited work.
9	Citing work partially confirms cited work.
10	Citing work is totally supported by cited work.
11	Citing work is partially supported by cited work.
12	Citing work is illustrated or clarified by cited work.
	Assumptive Type Citations
13	Citing work refers to assumed knowledge which is general background.
14	Citing work refers to assumed knowledge which is specific background
15	Citing work refers to assumed knowledge in an historical account.
16	Citing work acknowledges cited work pioneers

	Tentative Type Categories
17	Citing work refers to tentative knowledge.
	Methodological Type Categories
18	Use of materials, equipment, or tools.
19	Use of theoretical equation
20	Use of methods, procedures, and design to generate results
21	Use of conditions and precautions to obtain valid results
22	Use of analysis method on results
	Interpretational/Developmental Type Categories
23	Used for interpreting results.
24	Used for developing new hypothesis or model
25	Used for extending an existing hypothesis or model.
	Future Research Type Categories
26	Used in making suggestions of future research
	Use of Conceptual Material Type Categories
27	Use of definition
28	Use of numerical data.
	Contrastive Type Categories
29	Citing work contrasts between the current work and other work.
30	Citing work contrasts other works with each other.
	Reader Alert Type Categories
31	Citing work makes a perfunctory reference to cited work.32.
32	Citing work points out cited works as bibliographic leads
33	Citing work identifies eponymic concept or term of cited work
34	Citing work refers to more complete descriptions of data or raw sources of data.

2.2.3 Citation Sentiment Classification

Citation sentiment analysis, otherwise called polarity analysis, focuses on analyzing the texts around citations to classify citations into positive, neutral or negative sentiments. The difference between citation sentiment and polarity is that sentiments classes are finite-positive, neutral and negative, while polarity is between -1 and +1. Citation sentiment classifications are often based on polarity; citation contexts with polarity below zero are classified as negative sentiments, citation contexts with polarity above zero are classified as positive sentiments, and citation contexts with zero polarity are classified as neutral. Earlier studies on citation sentiment analysis have considered the sentence in which a citation occurs for citation sentiment analysis (Ritchie et al., 2008). Others studies recommended the *n* sentence window with the inclusion of *n* sentences or words before and *n* sentences or words after the citation sentence in sentiment analysis detection (Athar & Teufel, 2012b; Ritchie et al., 2008). However, Athar (2011), and Athar and Teufel (2012b) contended that the sentence-level sentiment analysis method might not work best for citation sentiment because of the complexity of sentiments expressed in scientific writings. Hedging, which is practised when expressing negative sentiments in academic writings, is an instance of the complexities in citation sentiment analysis. Athar (2011) therefore recommended that the n-gram and dependency relations work better for citation sentiment analysis. While (Xu et al., 2015) used the decision tree method with a sentence or paragraph window boundary because of the complexity of sentiments in the citation, Athar & Teufel, (2012a) considered a window of four sentences for citation context boundary marking.

Table 2.2: Citation Context Weighting/Classification Studies

Data Source	**Content (full-text, or sectional)** Qualitative Data						**Metadata** Quantitative Data			**…and External Data**
	CCA-Semantic Analysis			CCA-Syntactic Analysis			Statistical and Mathematical Analysis			
	Semantic similarity	Citation sentiment	Citation Function	Citation Location	Citation Frequency	Syntactic similarity	Metadata-based	Citation-count		
Metrics								PageRank	JIF	MI, co-citation
Types								Recursive	Non-recursive	
Mechanism/ based on	Semantic relationship btw whole or part (e.g. title, abstract, keywords, intro etc.) of the cited and citing papers.	Based on the polarity of citation context which could be positive, neutral or negative	The contribution of the cited paper in the citing paper which could be based on impact, the importance Perfunctory,	Location of the citation mention in the full-text	Number of citation mentions in the full-text	-Overlap btw references (bibliographic coupling) -Overlap btw keywords	-1/n of no reference -1/n of no authors -Publication age	Final value does not depend on the initial value; values a calculated recursively in such as way that highly influential citations are allocated more weights	The final value is equivalent to the initial value	Using the content or metadata and external data (data not found as the content or metadata of citing or cited the papers) such as the citation or publication number of the author as in MI or number of times the citing and cited papers have been co-cited

2.3 Citation Weighting

Citation weighting is concerned with allocating quantitative value or weights to citations. The task of allocating appropriate weight to citations is multidimensional as many citation weighting studies consider one or more dimensions such as size, impact, prestige, productivity, use, influence etc. Leydesdorff (2009) defined influence as a "combination of impact and productivity," while Prathap and Nishy (2016) and Prathap, Nishy and Savithri, (2016) defined influence as a combination of size (quantity) and impact (quality). Popularity was also defined by Yan and Ding, (2010) as the number of citations received by an article. Also, citation prestige was defined as the impact (which could be measured in impact factor) of the journal in which the citing article was published (Yan et al., 2011; Yan & Ding, 2010). In other words, more prestigious journals have higher impact factor while less prestigious journals have lower impact factor.

The review of the literature on citation weighting is classified based on the types of analysis studies on citation weighting perform. Ten dimensions of analyses in citation weighing were identified during the literature review of citation weighting studies (see Table 2.2). Citation weighting studies included the analysis of one or more elements of scientific publications in their methodologies. Citation location, for instance, has been included as one of or the only elements in citation weighting studies such as Hassan et al. (2018, 2017), Maričić, Spaventi, Pavičić, and Pifat-Mrzljak (1998), Pride and Knoth (2017), Strotmann and Zhao (2014).

2.3.1 Citation Weighting based on Semantic Similarity

Semantic similarity analysis is carried out in citation weighting studies to find out the relationship between the citing and cited articles. The semantic similarity analysis is employed on the premise that a cited document could only contribute to the citing document if there is a semantic relationship between the citing document and the cited document and that the closer the semantic relationship, the greater the influence of the cited document in the citing document. In other words, “the influence of a cited paper on a given citing paper (the citer) is proportional to the overlap in the semantic content of the cited paper and the citer” (Zhu et al., 2015, p. 412). Different methods and techniques have been employed for finding the semantic relationship between cited and citing articles for citation weighting purpose, mostly, by analyzing and extracting features and/or indexes from the full-texts, parts of the full-texts or metadata of the citing and cited articles.

Valenzuela, Ha, and Etzioni (2015) carried out a study on identifying important citations. One of the tested features is the semantic relationship between the abstract of the cited and citing papers using the td-idf cosine similarity scores. It was assumed that “the closer the abstracts, the more likely the new work extends the cited paper”. Thus, the extension of a cited work was marked as an indicator of importance. However, the study concluded that the relationship between the “degree of similarity between the abstracts of the cited and citing documnents” and citation influence was weak. Pride and Knoth, (2017), Hassan, Akram, and Haddawy (2017) and Hassan, Safder, Akram, and Kamiran (2018), as follow-ups to Valenzuela et al. (2015), also focused on the semantic relationship between the abstracts of the citing and cited papers by using the *td-idf* cosine similarity. According to

Pride and Knoth, (2017), the results of the study “demonstrate that abstract similarity between citing and cited paper is more predictive of citation influence” than previously shown because the abstract similarity measure provided the most important feature for identifying essential citations. Hassan et al. (2017) likewise concluded that the best performing feature for identifying important citations was the similarity between the abstract of the citing and the cited paper. These conclusions contradicted the position of Valenzuela et al. (2015).

While other studies on the relatedness between cited and citing documents have considered one of the sections of scientific publications, studies such as Zhu et al. (2015) have considered several parts of the citing and cited documents. Zhu et al. (2015) carried out a study on automatic citation weighting of the influence or importance of citations in the citing work. The study used title, abstract, introduction, conclusion and other core sections in the body of scientific articles, apart from the acknowledgements in measuring the semantic relationship. It was assumed that the title of a scientific article is a good summary or information surrogate about the full text. Another surrogate or summary of a full-text that was considered is the citation context, that is, from two words around the in-text citation to several sentences around it. The similarities between the citation contexts and the title, abstract, introduction and conclusion of the citing paper were calculated. The cosine similarity scores between the citation context and the abstract correlated most with academic influence, followed by context-conclusion, context-title, and context-introduction (Turney & Pantel, 2010)

While other studies focused on one or more sections of the full-texts or the metadata, Knoth and Herrmannova, (2014) considered the semantic similarity between full-texts of the cited and the citing papers for calculating contribution score, an indicator of citation weight. The contribution score was calculated by using the cosine similarity measure of *tf-idf* term-document vectors, as proposed by Manning, Raghavan, & Schütze (2009). The distance between the cited and citing papers was calculated as *dist*(d1, d2) = 1 — *sim*(d1, d2), where *sim*(d1, d2) is the cosine similarity of documents d1 and d2.

2.3.2 Citation Weighting based on Citation Function

Allocating weights based on citation function or motivation has mostly been done manually by allocating categorical weights. For instance, McCain and Turner, (1989) provided weights to citation functions with the weight of 0.5 to peripheral citations and 1.0 to central citations. Valenzuela, Ha, and Etzioni, (2015), focused on identifying essential citations as a binary classification task, providing a numeric scale as displayed in increasing order of importance as displayed in Table 2.3 below. Hassan, Akram, and Haddawy, (2017) extended the scheme in Table 2 by allocating the value of zero and one to the incidental and important citation classes respectively.

Table 2.3:Citation Annotation Labels

Citation Type	Fine-grained Label	Coarse Label
Related work	0	Incidental
Comparison	1	Incidental
Using the work	2	Important
Extending the work	3	Important

2.3.3 Citation Weighting based on Citation Location

Nazir et al. (2020) used Multiple Regression and Neural Network which are supervised machine learning models to allocate weights to citations based on their location in the surveyed scientific papers. As shown in Table 2.4, from the multiple regression analysis, the methodology section received the highest weight, followed by the results and discussions section, the introduction section, and the literature review section. Also, the normalized weights by neural networks allocated the highest weight to the citation in the results and discussions sections, followed by methodology, introduction, and literature review sections.

Table 2.4: Weights allocated by Regression Analysis in Nazir et al. (2020)

Sections	Weights	Weight Rank
Introduction	0.1891921316	3
Literature Review	0.1470393226	4
Methodology	0.3663496373	1
Results and Discussions	0.2974189085	2

2.3.4 Citation Context weighting based on Citation Frequency

Citation frequency or citation mention analysis deals with the number of times a citation is mentioned in a document. There are two types of citation frequency analysis; that is, the in-text citation frequency analysis and the reference analysis. In-text citation frequency analysis is the commonest in citation weighting and is based on the idea that cited publications that are mentioned more frequently are likely to contribute more to the citing document than those citations that are cited infrequently. Citation frequency remains the most adopted measure for citation weighing, and this shows its relevance to the subject.

Its main attraction is the simplistic nature of frequency analysis which is based on a simple or “rough” count of in-text citations (Zhao & Strotmann, 2014b).

Herlach, (1976) was one of earliest study that provided evidence for the use of citation frequency in quantifying the contribution of cited document in the citing document. Herlach (1976) considered the citation frequency analysis in finding the relationship between citing and cited documents for information retrieval by asking humans to rate the relevance of articles with multiple in-text citations and articles with single in-text citations to their cited publications. The result of the study indicated that articles with multiple in-text citations were found more relevant to the publications they cited than articles with single in-text citation by an approximate ratio of two to one. This provided a basis for other studies to assume that cited documents with more mentions contribute more than citations with less mention in the citing document. Voos & Dagaev, (1976) and Peritz, (1983), also studied the pattern of the distribution of citations in the sections of research papers. These studies concluded, like Herlach, (1976), that the citation frequency in a research paper is associated with the contribution of cited publication to the citing research paper.

Zhu, Turney, Lemire, and Vellino, (2015) is another study which sought to find out the most important research articles’ features for weighing citations. The research submitted that citation frequency was more important than other features such as citation location and semantics. Empirical studies such as (Boyack, van Eck, Colavizza, and Waltman, 2018; Ding, Liu, Guo, and Cronin, 2013; Strotmann and Zhao, 2014; (Pride and Knoth, 2017; Sánchez-Gil, Gorraiz, and Melero-Fuentes, 2018; Hu, Chen, and Liu, 2013 Hassan et al., 2017; Maričić, Spaventi, Pavičić, & Pifat-Mrzljak, 1998) have also shown that the citation frequency is important to citation weighing and that it provides a different indicator from the traditional citation count that could enhance research impact evaluation. Re-citation analysis is another method which has been proposed for citation frequency analysis by (Zhao & Strotmann, 2015). The re-citation approach considers only in-text citations that have been cited more than once while discounting uni-citations as perfunctory.

Zhao & Strotmann, (2016) compared 11 different citation weighting schemes based on the number of mentions for author ranking. These citation weighting schemes were either author or paper-based. Zhao and Strotmann, (2016), considered the following methods:

> *“Paper-Based Counting*
>
> 1. *cW1P is traditional citation counting, which adds 1 to an author’s citation count whenever a paper by this author is cited regardless of how many times this paper is cited there.*
>
> *2. cWnP adds N to an author’s citation count when a paper by this author is cited N times in a citing paper.*
>
> *3. cWn2P adds N2 to an author’s citation count when a paper by this author is cited N times in a citing paper.*
>
> *4. rW1P is a re-citation counting method that adds 1 to an author’s citation count for each paper by this author that is re-cited (i.e., cited at least twice) in a citing paper.*
>
> *5. rWnP adds N to an author’s citation count when a paper by this author is re-cited N times in a citing paper, that is, when it is cited N + 1 times there.*
>
> *6. rWn2P adds N2 to an author’s citation count when a paper by this author is re-cited N times in a citing paper.*
>
> *Author-Based Counting*
>
> *7. cW1A adds 1 to an author’s citation count if this author is cited in the text of a citing paper regardless of how many times this author is cited or how many papers by this author are cited there.*

8. *cWnA adds N to an author's citation count if this author is cited N times in the text of a citing paper regardless how many papers by this author are cited there; it always gives results identical to those of cWnP;*

9. *cWn2Aadds N2 to an author's citation count if this author is cited N times in the text of a citing paper regardless of how many papers by this author are cited there.*

10. *rW1A adds 1 to an author's citation count if this author is re-cited (i.e., cited at least twice) in the text of a citing paper regardless of how many times this author is cited or how many papers by this author are cited there.*

11. *rWnA adds N to an author's citation count if this author is re-cited N times in the text of a citing paper regardless of how many papers by this author are cited there."*

2.4 Citation Context in Citation Weighting

Citation context describes the expanse of words which have been written about a citation. Citation context was described by Khalid et al. (2018, p. 607) as "the text segments used to characterize a target citation" which could span a few words, phrases or sentences. Recognizing citation context is important to citation content, citation sentiment analysis, citation-based information retrieval and citation weighting research. Identifying citation context, like many human language problems, is a complex concept in computing because of the unstructured nature of human writings. While a citation context might span a phrase, another might span sentences or paragraphs. To address citation context problems in citation weighting studies, the two major methods that have been adopted are the rule of thumb and decision tree.

The commonest method which has been employed in citation context identification is the use of a fixed window or a specific number of characters, words or sentences in, before and after the citation sentence. The use of a fixed citation window is based on the assumption that citation contexts are captured in the text around the citation and therefore citation contexts could be identified by the marked boundaries. Doslu and Bingol (2016) specified “around 400 characters which are equally divided to both sides of citation marker” as their definition of citation context. Liu et al. (2014) defined the citation context as the sentence in which the cited publication is mentioned. Athar and Teufel (2012a, p. 598) showed some dynamism in applying the fixed window in that they considered “every sentence that is in a window of 4 sentences of the citation” for examination for references to the citation and sentences that did not contain a reference to the citation were excluded as citation context. Abu-Jbara, Ezra, and Radev (2013) also applied the fixed four sentence window, which only includes the sentence before the citing sentence, the citing sentence, and two sentences after the citing sentence. This method does not reflect the real-life citation context situation because of the dynamism of human languages. These citation contexts could be underrepresented or overrepresented.

Other studies have attempted to identify citation contexts without pre-setting a window, thereby identifying all the phrases or sentences that characterize a target citation. Khalid, Alam, and Ahmed (2018) proposed a heuristic algorithm which is built on the transition-based dependency parsing. Dependency parsers analyze the grammatical structure of sentences by capturing the syntactic relationship between the words in the sentence. Unlike other methods of citation context extraction, this method did focus on multiple and single reference text and subjective and objective citation contexts. This method appears to mirror the real-life situation of citation context identification as previous methods with fixed citation context windows have serious limitations.

2.5 Cascading Citation Analysis

Cascading citation is a relatively new research front for research assessment. It was proposed by Dervos and Kalkanis (2005) on the assumption that credits due to a publication should not be limited to direct citations (first-generation citation) it has received. The publication should also receive some form of credit for the citations (n-generations citations) to the publications that have cited it. Therefore, citations to an article cascade as it receives citations that are made "not just the number of citations made directly to the article in question, but also the ones made to the corresponding citing article(s)" (Dervos & Kalkanis, 2005, p. 668).

Articles are allocated $1/2^\wedge(n-1)$ citation from their nth generation citations. All the direct citations of an article are also categorized as its first-generation citations; in this case the value of n is 1. Therefore articles get a citation of 1 from all its direct citations. Subsequently, articles that cited the first-generation citations are the second-generation citations, and the original article is allocated a residual citation of $1\,/\,2$ (half) from the second-generation articles. All generations of citations apart from those that are categorized in the first-generation class are indirect citations. The original article receives ¼, 1/8,…1/2^{n-1} citation each from the third, fourth,…, and nth generations, respectively. Scholarly communications evaluations do not receive citation residual for second-generation citations, which have been claimed to be a necessity in Dervos and Kalkanis, (2005), Fragkiadaki, Evangelidis, Samaras, and Dervos (2010) because it is fair for an article to receive some residual citations from other articles that have cited the article that cited it.

According to Latour’s law of citation, a study is considered an unquestionable fact as the number of citations (influence) increases (Latour, 1987, Pg 51). Eventually, the idea may be reconstructed, shortened, eroded, distorted or become obliterated by incorporations to the extent that people will stop citing the original source. With the idea of cascading citation, there is a possibility that the influence of an article goes beyond its diminuendo (when people stop citing it) due to ageing/obsolescence (Burrel, 2001; Burrell, 2002, 2003). For instance, many studies mention and apply the “page rank” algorithm, an influential study, without citing Page et al. (1999), while its derivatives are cited. With the cascading citation weighting method, the original Page et al. (1999) will receive some indirect citations from the articles that cited the derivatives of the “page rank” algorithm.

It is perhaps reasonable to assume that the influence of an article may be theoretically infinitesimal because an article may have n generations citation, where n can be infinity. However, the publication may have significantly contributed and subsequently received few direct citations from very influential articles. These citing influential articles may have, in turn, received many direct and indirect citations. In this situation, the original article has received a fraction of its citations due to the current system of allocating credits for only direct citations. The cascading citation system presents a method that will potentially make it possible for studies to receive credit for the influence they have beyond the first generation.

2.6 Benchmark Datasets for Semantic Similarity

BIOSSES is a benchmark data set which is a collection of 100 sentence pairs from the biomedical domain, and the sentence pairs were selected from citation sentences. There are varying degrees of similarity between the citation sentence; some citation sentences cite the same reference articles for similar reason and are likely to be semantically similar while others cite different reference articles and are likely to be semantically different. The semantic similarity between the sentence pairs was determined manually by five different human experts, with the similarity scores set between 0 and 4. An evaluation of the performance of BioSentVec in determining semantic similarity between the BIOSSES sentence pairs gave a result of 0.795 correlation to the human annotation benchmark.

MedSTS is a dataset of 174,629 sentence pairs gathered from a clinical corpus of clinical notes at Mayo Clinic. A sample of 1250 sentence pairs were annotated by two medical experts with semantic similarity scores between zero and five (0=low to 5= high similarity). An evaluation of the performance of BioSentVec in determining semantic similarity between the MedSTS sentence pairs gave a result of 0.767 correlation to the human annotation benchmark.

2.7 Proposed Framework for Direct Citation Weighting

This thesis proposes a method for weighting citations, specifically those that are mentioned more than once in the citing publications, by placing a premium on the uniqueness of the citation contexts. This idea extends the citation mention analysis for weighting citations, where citation weights depend on the number of times a citation is mentioned in the citing document. The citation mention analysis assumes that all citation mentions are equal and allocate equal weights for each citation mentions. The whole citation context weighting research is built on analyzing citation contexts or citation mentions. This thesis is a paradigm shift in citation analysis because it is more interested in the contributions of the cited article than the quantitative values that are independent of these contributions.

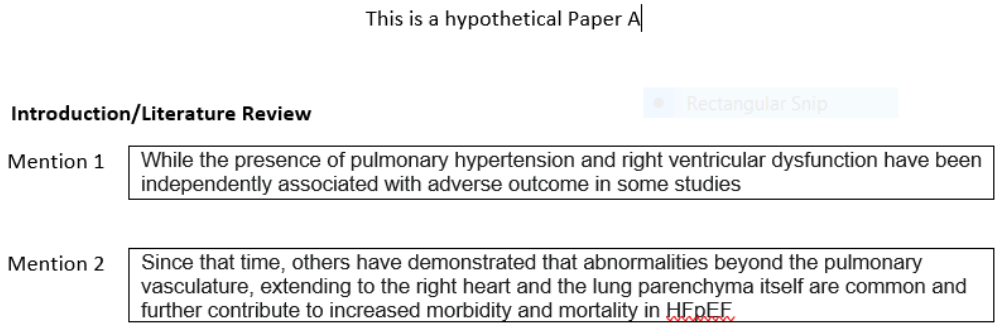
Identifying “contributions” in the citation contexts may be finding the semantic similarity between the *citation contexts* of the cited paper in the citing paper and can help to understand if the different citation mentions referenced different ideas from the cited paper. If all the citation contexts of a citation are the same or identical, the semantic similarity between citation contexts will be high or close to one. This study uses a semantic similarity score generated by a computer algorithm to determine the uniqueness of citation contexts. This is done by comparing the texts of the two citation contexts.

Citations were depicted with arrows that connect two citing and cited documents in citation networks. The contributions which are acknowledged through citations could only be analyzed through citation contexts. Though in citation networks, one point of connection exists between the cited and citing documents, this connection does not take into consideration the number of citation mentions of a citation. Each citation mention in a citing paper is a potential point of knowledge flow between the citing and cited the paper. Therefore the point of connection between citing and cited paper could be more than one in cases where there is more than one mention of a citation in the citing paper. For instance, citing document ‘E’ may cite the definition, the methodology and the results or conclusion of another document F with twelve in-text citations. The potential knowledge flows between the citing and cited paper or first-generation citations can be identified by analyzing the citation context.

2.7.1 Citation Classification using Semantic Similarity Scores

In practice, raw semantic similarity scores do not give precise information about texts that are *similar*, *somewhat similar* or *not similar*. By using inputs from human experts, boundaries that mark the three classes can be set. This framework starts with allocating weights to a citation that is mentioned twice. The framework for direct citation weighting centres around establishing the semantic similarity between multiple mentions of a citation. In order to execute a systematic strategy where the semantic similarity between all citation context pairs is possible, this framework recognizes the first citation mention in citations that are mentioned multiple times in the citing article as a unique citation context and is allocated the weight of one, regardless of the number of citation mentions. For instance, if a citation is mentioned twice (m_1, m_2) in a citing article, m_1 is automatically allocated the weight of one. Similarly, if a citation is mentioned three times (m_1, m_2, m_3) in a citing article, m_1 is also allocated a weight of one. This move is to establish a starting point for the comparison between the citation contexts.

The next stage is to find out the semantic similarity between the first and second citation contexts regardless of the number of citation mentions. In practice, the task is to find out if the second citation is different from the first. For instance, in Figure 2.1, after allocating the weight of one to *Mention 1*, the next move was to find the semantic similarity between *Mention 1* and *Mention 2*. If the semantic similarity between the two mentions is high and belongs to the *similar* semantic similarity class, it would be determined that the two are the same idea. In that case, *Mention 2* was adding nothing new to Mention 1, and zero was added to the weight of one that *Mention 1* had earlier, making one; the two citation contexts are identical. Otherwise, if the semantic similarity score was classified as *not similar*, the two citation contexts were counted as two different ideas. Therefore, one was added to the weight of *Mention 1*, making two; two different citation contexts.



This is a hypothetical Paper A

Introduction/Literature Review

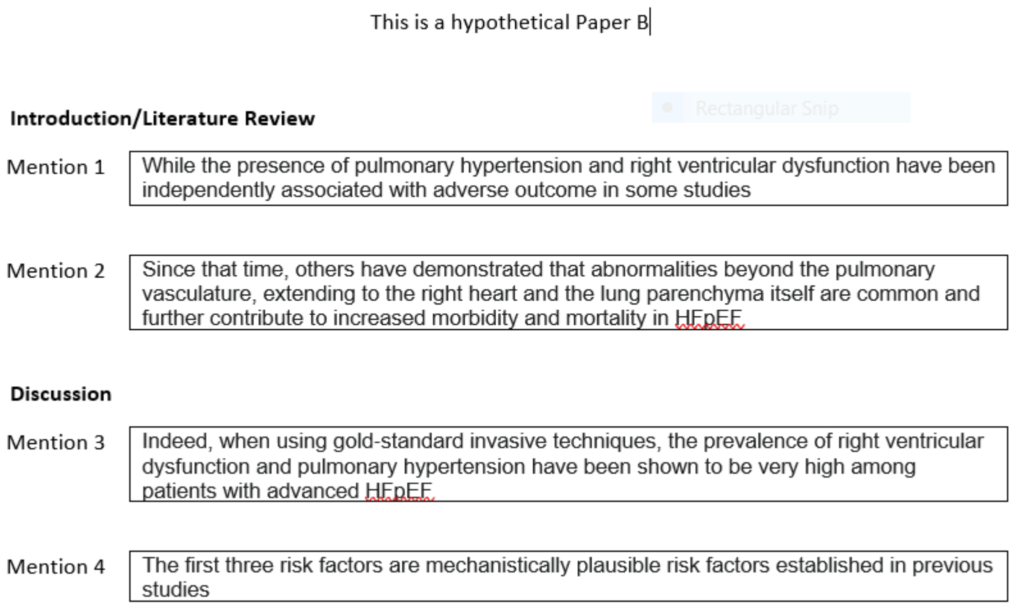
Mention 1 While the presence of pulmonary hypertension and right ventricular dysfunction have been independently associated with adverse outcome in some studies

Mention 2 Since that time, others have demonstrated that abnormalities beyond the pulmonary vasculature, extending to the right heart and the lung parenchyma itself are common and further contribute to increased morbidity and mortality in HFpEF

Figure 2.1: Paper A with Two mentions of Citation A

If the number of citation mentions is greater than two, the steps taken in the previous paragraph will precede the process. For instance, if there were four mentions of a cited article, as shown in Figure 2.2, the next step is to find out how the third mention-Mention 3- is different from the first two mentions -*Mention 1* and *Mention 2*. That is, having established the semantic similarity between *Mention 1* and *Mention 2*, and allocated weight to them accordingly, the next task was determining if *Mention 3* is different from *Mention 1* and *Mention 2*. This will be achieved by comparing *Mention 3* with *Mention 1* and *Mention 2* in pairs; *Mention 1*|*Mention 3* and *Mention 2*|*Mention 3*. The citation context pairs *Mention 1*|*Mention 3* and *Mention 2*|*Mention 3* were classified and weighted using their semantic similarity scores. For instance, if *Mention 1*|*Mention 3* were similar, a weight of zero was allocated, and if *Mention 2|Mention 3* were not similar, a weight of one was allocated. The next step was finding the average of the two weights and adding it to the weights that were obtained in the previous section.

The next step was to find out how Mention 4 was different from *Mention 1*, *Mention 2*, and *Mention 3* by comparing *Mention 4* to *Mention 3*, *mention 2*, and *Mention 1* in pairs; *Mention 1|Mention 4*, *Mention 2|Mention 4*, *Mention 3|Mention 4*. The next step was obtaining the semantic similarity scores of the three pairs (*Mention 1|Mention 4*, *Mention 2|Mention 4*, *Mention 3|Mention 4*), classifying the citation context pairs based on their semantic similarity scores, and allocating weights to them. The last step was adding the average of the weights to what was obtained previously



This is a hypothetical Paper B

Introduction/Literature Review

Mention 1 — While the presence of pulmonary hypertension and right ventricular dysfunction have been independently associated with adverse outcome in some studies

Mention 2 — Since that time, others have demonstrated that abnormalities beyond the pulmonary vasculature, extending to the right heart and the lung parenchyma itself are common and further contribute to increased morbidity and mortality in HFpEF

Discussion

Mention 3 — Indeed, when using gold-standard invasive techniques, the prevalence of right ventricular dysfunction and pulmonary hypertension have been shown to be very high among patients with advanced HFpEF

Mention 4 — The first three risk factors are mechanistically plausible risk factors established in previous studies

Figure 2.2: Citation with Four Mentions

This framework was proposed to complement the citation mention analysis, which counts the number of times a citation is mentioned in a paper. As opposed to counting the number of mentions in a paper alone, this study proposed considering the number of unique citation contexts that have been referenced by the multiple mentions of a citation. Multiple ideas from a cited paper may have been mentioned at different locations of the citing paper and should be allocated more weights. Otherwise, identical ideas should attract lower weights.

This proposed framework has two merits: first, the interaction between every citation mention was captured in the comparisons, regardless of the number of citation mentions. N combination 2 ($^nC_2$) number of citation context pairs were obtained, where n=the number of citation mentions. Second, each citation mention had a possibility of being assigned a maximum weight of one.

2.8 Proposed Framework for Residual Citation Weighting

Allocating weights to indirect citation mentions is different from allocating weights to direct citations. Direct citations weighting was done for multiple citation mentions in a citing document. On the other hand, 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 n+1th generation article. Citation contexts are surrogates of cited publications' contribution.

On a citation network, citation paths or citation chains exist 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 determine if 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 2.3, it can be observed that 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.

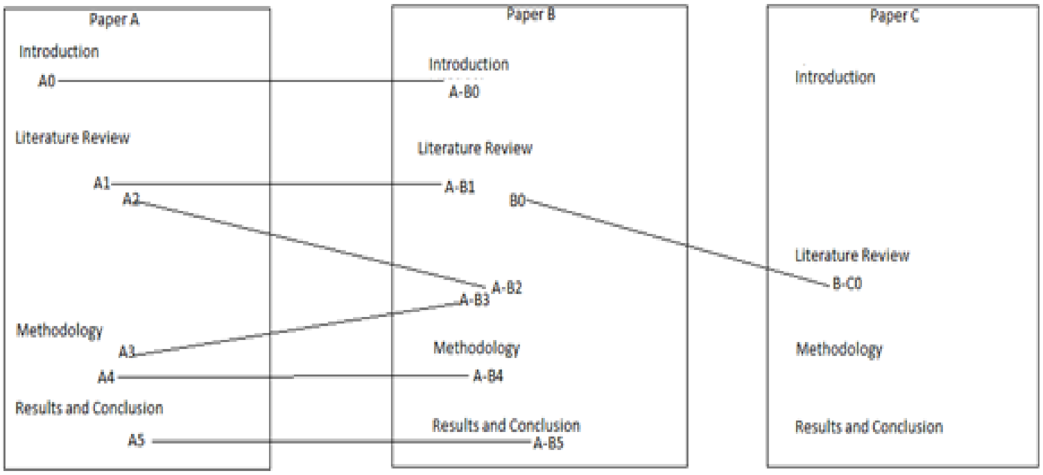


Figure 2.3: Three publications on a citation chain

Importantly, the theoretical framework assumes that the citation context pairs with the highest semantic similarity can be used to assess the weight of residual citation that should accrue to the base article. Using the above example, the citation context pairs with the highest semantic similarity 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.

2.9 Summary of the Literature Review Chapter

The literature review chapter is divided into eight sections. The first section presents a literature review of citation context analysis; this provides a background review of citation context literature. A review of literature on citation classification was presented in the second section. Research literature covering classifications such as citation function, citation sentiment and citation importance were explored. The third section of the literature review covered citation weighting analytical methods; these methods were reviewed based on semantic similarity, citation function, citation location and citation frequency. The fourth section presents a review of the literature on the significance of citation context in citation weighting studies. The fifth section focused on the cascading citation system. The seventh and eight sections contain proposed frameworks for the direct and residual citation aspects of this study, respectively.

Chapter 3

3 Methodology

The methodology for this proposed study is written in seven sections. The data sampling strategy is presented in section one. Section two provides details on the data collection strategies and the steps that were taken to extract citation context from the articles' full text. Data pre-processing is presented in section three with details of the strings that were either removed or replaced so that the semantics of the citation context were not jeopardized. The data annotation by experts was explained in section four. The citation similarity algorithm which is based on cosine distance between two citation contexts, is presented in section five. The citation weighting method, which is based on the citation context similarity, is presented in section six. Data analysis methods for the direct and indirect citation aspects of this thesis are provided in sections six and seven, respectively.

3.1 Data Sampling

The two datasets for this thesis were collected from the PubMed database, using the Web of Science search interface, since there are no available annotated datasets for cascading citation scheme that incorporate citation weighting. The scope of the study is limited to the biomedical sciences disciplines because of the readily available robust computational knowledge representation models such as word/sentence vectors space models and word embeddings, which were scarcely available for other scientific disciplines.

3.1.1 Sampling Articles for Direct Citation Context Weighting

The first dataset meant for direct citation consists of one hundred (100) articles sampled from moderately cited articles published in 2014 and indexed in MEDLINE. Moderately cited articles were operationalized as articles that received 50 to 200 citations within five years of publication, that is, between 2014 and 2019. To retrieve the biomedical publications that were published five years before 2019 from the WoS, MEDLINE was searched with the "year published" set at 2014. The query returned 1,037,524 results, and when the results were limited to journal articles alone, the numbers reduced to 950,143. A sampling frame was created from articles that had 200 to 50 citations after sorting the articles in decreasing number of citations. 58,938 fell within this category and were ranked between 5,230th and 64,168th. Every 590th article was sampled: 5,230, 5,820, 6,410….63,640th articles. The 100 articles dataset for direct citation was collected between the 11th and 23rd January 2020.

3.2.2 Sampling Articles for Indirect Citation Weighting

The process of sampling articles for the indirect citation aspect of this study was quite different from that of the direct citation. Sampling, in this case, was done in six stages with the first stage for the base articles, and the last five stages for five generations of citations, one stage per generation. The ten base articles were sampled as the ten most-cited biomedical articles that were published in the year 2014. Articles in other fields such as Materials Science that ranked among the most cited biomedical articles were ignored and replaced with articles that ranked next to them because after probing down from their first to subsequent generations, their citation bases shifted more from biomedical fields to their core fields. Example of such articles include Zhou, et al. (2014)[1]. This step was taken because the semantic similarity algorithm was based on a knowledge representation model that was trained on biomedical publications. Therefore, including articles outside the domain of the training corpus could produce non-optimal result in the semantic similarity tasks. Duplicates of already sampled articles were also not included.

For the first stage 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.

[1] “Zhou, et al. (2014) *Photovoltaics. Interface engineering of highly efficient perovskite solar cells.* Science Vol. 345 (6193).”

Duplicates were not recorded during the data collection, i.e., once an article is sampled, it was not sampled again in subsequent generations. For instance, Ross et al (2016)[2] and Alexander et al (2017)[3] were second-generation citations for one of the base articles, Alexander et al (2017) also cited the Ross et al (2016), and ranked as one of its two most cited articles citing articles. However, Alexander (2017) was not sampled as a third-generation article. 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)[4] 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. Other versions of identical articles were

[2] Ross et al (2016) *2016 American Thyroid Association Guidelines for Diagnosis and Management of Hyperthyroidism and Other Causes of Thyrotoxicosis*. Thyroid 26(10) Pp1343-1421”

[3] Alexander et al (2017) *2017 Guidelines of the American Thyroid Association for the Diagnosis and Management of Thyroid Disease During Pregnancy and the Postpartum”* Thyroid 27(3) Pp.315-389

[4] Krittanawong C. et.al. (2019) *Deep learning for cardiovascular medicine: a practical primer”* European Heart Journal 40 (25). Pp 2058–2073

excluded after the inclusion of the most cited. For instance, Aboyans et. al (2017)[5] had three other duplicates in the WoS database. Articles that mentioned a cited paper more than 20 times in the full text were ignored for ease of computation. For example, Hendrickx et al., (2019) was excluded because of 63 citation mentions of Sidor & Hopson (2018).

The sampling was executed between November 22, 2019 and January 11, 2020 to produce a sampling frame, but selected articles in the first three generations, where full texts were not available were replaced. Unavailable articles in the fourth and fifth generations were not replaced because at the fourth generation some articles stopped having indirect citation.

3.2 Data Collection

Data collection for the selected articles for direct and indirect citation aspects of this study proceeded in two parts. First, full texts of the citing papers in the two aspects were downloaded so that citation contexts could be extracted. At this stage of the study, unavailable full texts were not included in the direct citation aspect of the study. Similarly, some articles were excluded because the cited articles were not found in the full texts or

[5] Aboyans V, et. al (2017). ESC Guidelines on the Diagnosis and Treatment of Peripheral Arterial Diseases, in collaboration with the European Society for Vascular Surgery (ESVS): Document covering atherosclerotic disease of extracranial carotid and vertebral, mesenteric, renal, upper and lower extremity arteries. Endorsed by: the European Stroke Organization (ESO)The Task Force for the Diagnosis and Treatment of Peripheral Arterial Diseases of the European Society of Cardiology (ESC) and of the European Society for Vascular Surgery (ESVS). Eur Heart J 2017;39:763–816.

the only citation(s) were part of Table or Figure. Therefore, the citation count for each of the 100 direct citation articles reduced by a little bit. For the indirect citations, a sampled article, whose full text was unavailable at the time of data collection, was replaced with an available article that ranked next in the number of citations.

For the direct citation aspect of the study, citation contexts of cited scientific papers in the citing papers were extracted. Whereas the citation contexts of the ten base articles from the full texts of all the five generations of publications were extracted, data collection was done manually by identifying and extracting the span of texts that accurately represent the context of idea that was referenced from an in-text citation. Citation context can span a few words within a sentence (a phrase), one or more sentences or a paragraph. The sentences that accurately represent the citation context are usually the texts around the in-text references, while some other times, citation contexts are located separately from the in-text citation.

The task of identifying citation contexts is not a trivial manual or computing task. Studies that have adopted the automatic means of identifying/extracting citation contexts have used a fixed window of sentences or words before and/or after the in-text references (Caragea et al., 2014; Dong & Schafer, 2011; Houngbo & Mercer, 2017; Singha Roy et al., 2020). While this method looks easy and fast, cleaning and preprocessing full texts of scientific papers to machine-readable formats is not trivial, considering the task that is involved in converting the different files and referencing formats into a format that is useable for NLP/computation. Besides, using the fixed window of words or sentences is prone to errors; citation contexts are misrepresented in cases where citation contexts are not the texts around the in-text references, and they are under/over-represented in cases where the span of text is shorter or longer than the fixed window. Citation contexts were identified manually in this study, though this method was more cumbersome than the use of a fixed window. This path was chosen to collect more accurate citation contexts.

The citation context was the span of texts about the cited publication's in-text citation that explains or represents the contribution referenced by the citing article. Most times, the citation contexts were located as the texts around the in-text citations and a few times, the citation contexts occurred in words after or before the in-text references. The citation contexts were collected from the publications that cited the 100 sampled articles for the direct citation aspect of the study. In the case of the indirect citations, the citation contexts of the ten base articles and those of first to fifth generation publications were collected.

Rule that were observed are discussed in the next sub sections. Citation context examples are in italics.

3.2.1 Single Sentence Citation Contexts

This is the commonest citation context and most of the citation contexts belong to this category. In this case, the sentence in which the citation is mentioned adequately represents the citation context.

3.2.2 Continuous Multiple Sentence Citation Contexts

Most of the multiple sentence citation contexts belong to this category as citation contexts span the sentence in which the citation is mentioned and the sentence(s) before and/or after.

3.2.3 Multiple Sentence Citation Contexts with implicit references

Sometimes, multiple sentence citation contexts are represented with sentences that do not follow each other sequentially. In this case, the citation sentence contains an implicit reference to a list or object which is mentioned in the sentence(s) that are not placed directly before or after it.

Considering an in-text citation referred to as 11, this citation sentence "The symptoms appear to be more frequent in physically active individuals.[11]" makes an implicit reference to "The symptoms". "The symptoms" were listed in the first sentence of the paragraph as shown below:

> *In the absence of a standardized classification of SAMS, we propose to integrate all muscle-related complaints (e.g. pain, weakness, or cramps) as 'muscle symptoms', subdivided by the presence or absence of CK elevation (Table 1). Pain and weakness in typical SAMS are usually symmetrical and proximal, and generally affect large muscle groups including the thighs, buttocks, calves, and back muscles. Discomfort and weakness typically occur early (within 4–6 weeks after starting statin therapy[22]), but may still occur after many years of treatment. Onset of new symptoms may occur with an increase in statin dose or initiation of an interacting drug. The symptoms appear to be more frequent in physically active individuals.[11] Statin-associated muscle symptoms often appear more promptly when patients are re-exposed to the same statin.*

Therefore, this sentence *"In the absence of a standardized classification of SAMS, we propose to integrate all muscle-related complaints (e.g. pain, weakness, or cramps) as 'muscle symptoms…"* was included as part of the citation context for citation 11

3.2.4 Multiple unique in-text citations in a sentence

This occurs when more than one in-text citation is mentioned in a sentence. In each of the instances below, two contexts were extracted.

<u>Instance 1</u>: *"We used the GBD 2013 results for YLLs[2] and YLDs[1] to calculate DALYs."*

Context 1 :*"We used the GBD 2013 results for YLDs[1] to calculate DALYs."*

Context 2:*"We used the GBD 2013 results for YLLs[2] to calculate DALYs"* respectively.

<u>Instance 2</u>: *Wang and colleagues[2] have described data sources and methods to estimate mortality and life tables, and Vos and colleagues[1] have described these for the measurement of prevalence of sequelae and disability weights.[1]*

Context 1: *Wang and colleagues[2] have described data sources and methods to estimate mortality and life tables.*

Context 2: *Vos and colleagues[1] have described these for the measurement of prevalence of sequelae and disability weights.[1]*

<u>Instance 3</u>: *The notion of the epidemiological transition has been expanded to recognise the phase in transition that leads to double burden of disease9, 28, 29 and the* *countertransitions of the HIV/AIDS epidemic and the rise of mortality in the former Soviet Union.*2, 10, 11, 13, 30, 31, 32.

Context 1: *(in-text citations 9, 28, 29): The notion of the epidemiological transition has been expanded to recognise the phase in transition that leads to double burden of disease and the rise of mortality in the former Soviet*

Context 2: *(in-text citations 2, 10, 11, 13, 30, 31, 32): The notion of the epidemiological transition has been expanded to recognise the phase in transition that leads to the countertransitions of the HIV/AIDS epidemic and the rise of mortality in the former Soviet*.

3.2.5 Multiple mentions of a citation in a sentence

When an in-text citation appears more than once in a sentence, the two mentions are treated as one. For instance, citation 1 was mentioned twice in *"Wang and colleagues*[2] *have described data sources and methods to estimate mortality and life tables, and Vos and colleagues*[1] *have described these for the measurement of prevalence of sequelae and disability weights.*[1]*"*.

The in-text citation 1 has only one context: *"Vos and colleagues*[1] *have described these for the measurement of prevalence of sequelae and disability weights.*[1]*"*

3.2.6 Citation context phrases

In some instances, including all the texts in the citation sentence may not accurately represent the citation context as the citation context may only be represented by a phrase, which is less than a whole sentence. For instance, in this sentence "*Muscle pain or aching,* *stiffness, tenderness or cramp (often referred to as 'myalgia'*[19]*) attributed by patients to their statin use is usually symmetrical but may be localized, and can be accompanied by muscle weakness; any of these effects occur predominantly without an elevation of CK"* the context of citation 19 is *"Muscle pain or aching, stiffness, tenderness or cramp (often referred to as 'myalgia')"*

3.2.7 In-text citation that appears in two or more consecutive sentences

In cases where an in-text citation is mentioned in sentences following one another and references or mentions of at least a keyword in one of the two sentences is made in the second sentence, the two in-text citations were treated as a citation context, regardless of the relationships between the citation contexts. Example is in-text citation 20 in this sentence: *Few other RCTs have queried for muscle complaints among participants.*[20] *Muscle complaints in other clinical trials have been similar in statin-treated and placebo subjects.*[4,20,23,24]. Citation context for in-text citation becomes: *Few other RCTs have queried for muscle complaints among participants. Muscle complaints in other clinical trials have been similar in statin-treated and placebo subjects*.

3.2.8 References to Tables, Figures, Appendices and Supplementary Materials

In-text citations in and references to Tables, Figures, Appendices supplementary materials and their headings were ignored. For example, the reference to supplementary material in this context was ignored: *Indeed, a definitive diagnosis of SAMS is difficult because symptoms are subjective and there is no 'gold standard' diagnostic test. Importantly, there is also no* *validated muscle symptom questionnaire, although the National Lipid Association has proposed a symptom scoring system based on the STOMP trial and the PRIMO survey (see Supplementary material online, Table S2)*

3.2.9 List explicitly, to uncover hidden in-text citation

For Example, this citation context: *In randomized, controlled trials (RCTs), adverse event rates (including complaints of muscle pain) are similar in statin and placebo groups,*[2–4]

Becomes: *In randomized, controlled trials (RCTs), adverse event rates (including complaints of muscle pain) are similar in statin and placebo groups,*[2, 3, 4] so that in-text citation context 3 will explicitly be listed as an in-text citation.

3.2.10 In-text citations in lists

Below are in-text citations 28, 29 and 30.

Pressure and flow tracings in the last minute at each phase were analyzed, and the following parameters were collected:

(1) The mean P_{aw} during either the inspiratory or expiratory phase[28]*;*

(2) The peak inspiratory and expiratory flow rate (PIF and PEF);

(3) The inspiratory V_T integrated by flow tracing, and RR and minute ventilation (MV);

(4) The P_{es} swing during inspiration (ΔP_{es})[29,30]*;*

Becomes two contexts:

Context 1 (for in-text citation 28): *Pressure and flow tracings in the last minute at each phase were analyzed, and the following parameters were collected: The mean P_{aw} during either the inspiratory or expiratory phase;*

Context 2 (for in-text citation 29 and 30): *Pressure and flow tracings in the last minute at each phase were analyzed, and the following parameters were collected: The peak inspiratory and expiratory flow rate (PIF and PEF);The inspiratory* V_T *integrated by flow tracing, and RR and minute ventilation (MV); The* P_{es} *swing during inspiration (*ΔP_{es}*);*

3.2.11 Exceptions

Some citation mentions were ignored because they did not represent citation contexts of the cited article.

Instance 1: Some editorials that announce or promote articles in issues of a journal sometimes include in-text citations that do not represent knowledge from the cited document. Some of the citation mentions in such editorials were not recorded.

Example (in-text citation 1) [6]:

We thank for Dr Čulić's interest in our article,[1] *and his comment on the public health relevance of our findings*

Instance 2: Citation mentions as part of Tables, Figures and Supplementary materials. Citation mentions in equations.

[6] Chen, K., Peters, A., Schneider, A., Peters, A., Schulz, H., Schwettmann, L., Leidl, R., Heier, M., & Strauch, K. (2019). Burden of myocardial infarctions attributable to heat and cold. *European Heart Journal*, *40*(41), 3440–3441. https://doi.org/10.1093/eurheartj/ehz612

Instance 3: Unusually long (typically more than one paragraph) in-text citations were also not included. Biosent2vec, the computer algorithm for semantic similarity measurement that was used for this study works better on sentences, supplying citation contexts that span paragraphs may produce sub-optimal results. This is a limitation of this study; however citation contexts in this category are very rare. An example is the citation context below.

On the basis of NHANES 2011 to 2012, the average dietary consumption by US children and teenagers of selected foods and nutrients related to cardiometabolic health is detailed below[3]:

- *Whole grain consumption was <1 serving per day in all age and sex groups, with <5% of all children in different age and sex subgroups meeting guidelines of ≥3 servings per day.[17]*
- *Fruit consumption was low and decreased with age: 1.7 to 1.9 servings per day in younger boys and girls (5–9 years of age), 1.4 servings per day in adolescent boys and girls (10–14 years of age), and 0.9 to 1.3 servings per day in teenage boys and girls (15–19 years of age). The proportion meeting guidelines of ≥2 cups per day was also low and decreased with age: ≈8% to 14% in those 5 to 9 years of age, 3% to 8% in those 10 to 14 years of age, and 5% to 6% in those 15 to 19 years of age. When 100% fruit juices were included, the number of servings consumed increased by ≈50%, and proportions consuming ≥2 cups per day increased to nearly 25% of those 5 to 9 years of age, 20% of those 10 to 14 years of age, and 15% of those 15 to 19 years of age.*
- *Nonstarchy vegetable consumption was low, ranging from 1.1 to 1.5 servings per day, with <1.5% of children in different age and sex subgroups meeting guidelines of ≥2.5 cups per day.*
- *Consumption of fish and shellfish was low, ranging between 0.3 and 1.0 servings per week in all age and sex groups. Among all ages, only 7% to 14% of youths consumed ≥2 servings per week.*
- *Consumption of nuts, seeds, and beans ranged from 1.1 to 2.7 servings per week among different age and sex groups, and generally <15% of children in different age and sex subgroups consumed ≥4 servings per week.*

- *Consumption of unprocessed red meats was higher in boys than in girls and increased with age, up to 3.6 and 2.5 servings per week in 15- to 19-year-old boys and girls, respectively.*

- *Consumption of processed meats ranged from 1.4 to 2.3 servings per week, and the majority of children consumed <2 servings per week of processed meats.*

- *Consumption of SSBs was higher in boys than in girls in the 5- to 9-year-old (7.7±6.2 versus 6.0±3.8 servings per week) and 10- to 14-year-old (11.6±5.3 versus 9.7±7.9 servings per week) groups, but it was higher in girls than in boys in the 15- to 19-year-old group (14±6.0 versus 12.4±5.8 servings per week). Only about half of children 5 to 9 years of age and one-quarter of boys 15 to 19 years of age consumed <4.5 servings per week.*

- *Consumption of sweets and bakery desserts was higher among 5- to 9-year-old and 10- to 14-year-old boys and girls and modestly lower (4.7 to 6 servings per week) among 15- to 19-year-olds. A minority of children in all age and sex subgroups consumed <2.5 servings per week.*

- *Consumption of eicosapentaenoic acid and docosahexaenoic acid was low, ranging from 0.034 to 0.065 g/d in boys and girls in all age groups. Fewer than 7% of children and teenagers at any age consumed ≥0.250 g/d.*

- *Consumption of SFAs was ≈11% of calories in boys and girls in all age groups, and average consumption of dietary cholesterol ranged from ≈210 to 270 mg/d, increasing with age. Approximately 25% to 40% of youths consumed <10% energy from SFAs, and ≈70% to 80% consumed <300 mg of dietary cholesterol per day.*

- *Consumption of dietary fiber ranged from ≈14 to 16 g/d. Fewer than 3% of children in all age and sex subgroups consumed ≥28 g/d.*

- *Consumption of sodium ranged from 3.1 to 3.5 g/d. Only 2% to 11% of children in different age and sex subgroups consumed <2.3 g/d.*

3.3 Data Pre-processing

Extracted citation contexts were pre-processed for computing so that unwanted texts would not interfere in the linguistic features of the dataset. The following steps were taken for pre-processing.

3.3.1 Implicit and Explicit in-text citations

Implicit and explicit in-text citations were treated differently. While implicit in-text citations were removed, explicit in-text citations were replaced. Implicit in-text citations example is (Sergio, 2020) or (Sergio et.al., 2020), while on the other hand, Sergio (2020) or Sergio et. al. (2020) is an example of explicit citation.

As part of the preprocessing, place holders for implicit in-text citations were removed. Placeholders for implicit in-text citations refer to the texts that represent the in-text citation, and depending on the referencing format, could be numerals or alphanumeric (e.g. (Sergio, 2020), [25-90], 23-34). The placeholders for implicit in-text references add no semantic value to the citation contexts; rather their presence may unnecessarily alter the semantics of the citation context.

For instance, 23, 24 and 92-94 were removed from the two citation contexts below.

Citation context 1: *Some studies examine associations with income per person, whereas others use variables such as mean age of the population 23, 24.*

Citation context 2: *The interaction of statins with muscle mitochondria can involve (i) reduced production of prenylated proteins including the mitochondrial electron transport chain (ETC) protein, ubiquinone (coenzyme Q10), (ii) subnormal levels of farnesyl pyrophosphate and geranylgeranyl pyrophosphate leading to impaired cell growth and autophagy, (iii) low membrane cholesterol content affecting membrane fluidity and ion channels, and (iv) the triggered calcium release from the sarcoplasmic reticulum via ryanodine receptors, resulting in impaired calcium signalling.*[92–94]

Explicit in-text citations were replaced with in_text_ref. For example: *Our findings support those of Salomon and colleagues,*[8] *which showed that HALE is increasing more slowly than life expectancy: ie, as life expectancy increases, the expectation of years lived with multiple sequelae increases as well.* became *Our findings support those of in_text_ref which showed that HALE is increasing more slowly than life expectancy: ie, as life expectancy increases, the expectation of years lived with multiple sequelae increases as well.*

3.3.2 Acronyms

Ambiguous and unambiguous acronyms were treated differently. Ambiguous acronyms were replaced with their full meanings while unambiguous acronyms were not replaced. Ambiguous acronyms refer to acronyms that could have many meanings at different places or contexts, in biomedical or general texts. Disambiguation, with the aim of increasing precision, makes it necessary to replace the ambiguous acronyms with full meanings. Examples of ambiguous acronyms with different full meanings include commonly used medical terms such as ER, which does not always mean emergency room; but it could also mean endoplasmic reticulum. Another ambiguous acronym is CV is often construed as curriculum vitae, but could also mean cardiovascular or cardinal vein. Other examples of ambiguous acronyms that have different full meanings in different countries or contexts include EPA, which could stand for Evolutionary Placement Algorithm, Environmental Protection Agency (United States of America), European Psychiatric Association (Europe), Environmental Protection Authority (Australia), Environmental Protection Act (Canada) or Eicosapentaenoic Acid. Another example of ambiguous acronym is ML methods which could mean machine learning methods or maximum likelihood methods. ICD could mean implantable cardioverter–defibrillator or International Classification of Diseases or induced circular dichroism. Some publications such as editorials are written colloquially and acronyms are sometimes written without definitions in full the first time they are mentioned. This makes automatic disambiguation difficult simply through referencing or dependency relations.

Another set of ambiguous acronyms are represented by tokens that also amount to English words. Examples are WHO which is an English word-*who* (a pronoun) and also an acronym for the World Health Organization; NO could either be “nitric oxide”, “neuromyelitis optica” or “*no”* as in opposite of yes. US could be United States, *us* (a pronous), or ultra-sound. ACE could stand for Adverse Childhood Experiences, Angiotensin-converting enzyme, or the word *ace* (an expert or champion).

While it is necessary to adequately disambiguate, some acronyms could lose their meanings if they are replaced, and therefore unambiguous acronyms were not replaced with their full meanings. For instance, replacing HIV with its full meaning could make it lose its value because HIV has been the generally accepted name for the human immunodeficiency viruses which cause a disease called acquired immunodeficiency syndrome (AIDS). Therefore, HIV represents the disease more in texts than its full meaning. Other common examples of unambiguous acronyms include drug names (e.g. NAMI-A and KP1019, CRISPR), DNA/RNA (e.g. LncRNA SNHG3), computer tools (e.g. FastQC, RDP4) and research areas (e.g. GWAS)

Another side of disambiguation is collocation. Full meaning of acronyms also replaced acronyms where they were found to be used inconsistently. For instance, Resveratrol was denoted as RES in (Liao et al., 2018)[7] while it was denoted as RSV in (Annunziata et al., 2019)[8].

3.3.3 Direct Citation Context Pairing

This stage of data pre-processing proceeded after the previously described pre-processing steps. Citation context pairing was done differently for the direct and indirect citation contexts categories. For the 100 sampled articles in the direct citation aspects of this study, only citing articles that referenced the sampled articles at least twice in their full texts were included in the citation context pairing. Citation context pairing occurred between the mentions of a cited document in the citing document. Therefore, if a cited paper is mentioned twice in the citing paper, the number of citation context pairs will be one. If a

[7] Liao, W., Yin, X., Li, Q., Zhang, H., Liu, Z., Zheng, X., Zheng, L., & Feng, X. (2018). Resveratrol-Induced White Adipose Tissue Browning in Obese Mice by Remodeling Fecal Microbiota. *Molecules*, *23*(12). https://doi.org/10.3390/molecules23123356

[8] Annunziata, G., Maisto, M., Schisano, C., Ciampaglia, R., Narciso, V., Tenore, G. C., & Novellino, E. (2019). Effects of Grape Pomace Polyphenolic Extract (Taurisolo®) in Reducing TMAO Serum Levels in Humans: Preliminary Results from a Randomized, Placebo-Controlled, Cross-Over Study. *Nutrients*, *11*(1). https://doi.org/10.3390/nu11010139

cited paper is mentioned four times, the number of citation context pairs will be six. Number of citation context pairs was *n* combination 2 ($^{n}C_2$), where *n* is the number of citation mentions.

Using | to denote pairing of the two contexts, therefore, for a citation that is mentioned twice, M_a and M_b, the citation context pairs are $M_a|M_b$. For a citation that is mentioned three times M_a, M_b, M_c, the citation context pairs are $M_a|M_b$, $M_a|M_c$ and $M_b|M_c$.

For a citation that is mentioned four times M_a, M_b, M_c, M_d the citation context pairs are $M_a|M_b$, $M_a|M_c$, $M_b|M_c$, $M_a|M_d$, $M_b|M_d$, and $M_c|M_d$.

For a citation that is mentioned five times M_a, M_b, M_c, M_d, M_e the citation context pairs are $M_a|M_b$, $M_a|M_c$, $M_b|M_c$, $M_a|M_d$, $M_b|M_d$, $M_c|M_d$, $M_a|M_e$, $M_b|M_e$, $M_c|M_e$, and $M_d|M_e$.

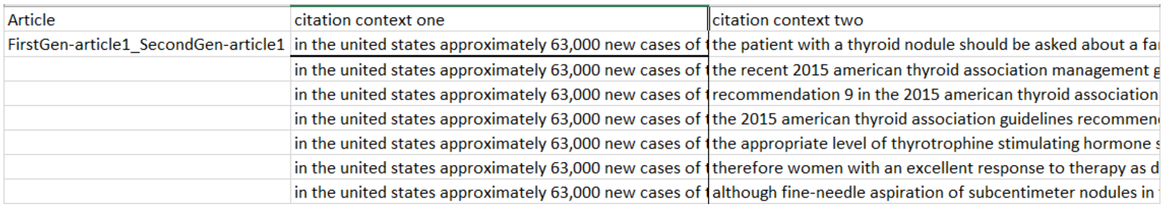
For a citation that is mentioned six times M_a, M_b, M_c, M_d, M_e, M_f the citation context pairs are $M_a|M_b$, $M_a|M_c$, $M_b|M_c$, $M_a|M_d$, $M_b|M_d$, $M_c|M_d$, $M_a|M_e$, $M_b|M_e$, $M_c|M_e$, $M_d|M_e$, $M_a|M_f$, $M_b|M_f$, $M_c|M_f$, $M_d|M_f$ and $M_e|M_f$.

3.3.4 Indirect Citation Context Pairing

An indirect citation context pair refers to two points on a citation chain. The first point is the mention of the base article in the first generation article, and the second point is the mention of the nth generation article in the n+1th generation article. For simplicity's sake, let us assume that the base article was mentioned three times in the first generation citation. Then, we can represent the three citation contexts as $M_{a\text{-}01}$, $M_{b\text{-}01}$, and $M_{c\text{-}01}$. Using the same convention, and assuming that the 1st generation citation was mentioned four times in the

2nd generation citations four times, we can represent the citation contexts of the 1st generation citation in the second-generation citation as M_{a-12}, M_{b-12}, M_{c-12}, and M_{d-12}. Then the citation context pairs will be $M_{a-01}|M_{a-12}$, $M_{b-01}|M_{a-12}$, $M_{c-01}|M_{a-12}$, $M_{a-01}|M_{b-12}$, $M_{b-01}|M_{b-12}$, $M_{c-01}|M_{b-12}$, $M_{a-01}|M_{c-12}$, $M_{b-01}|M_{c-12}$, $M_{c-01}|M_{c-12}$, $M_{a-01}|M_{d-12}$, $M_{b-01}|M_{d-12}$, $M_{c-01}|M_{d-12}$

So, in general, if there are *n* number of citation contexts of a base article in first generation articles, and *m* number of citation contexts of the first-generation article in the second generation article, the number of pairs obtainable for the first-second generation citation context comparison is *n* multiplied by *m*. For instance, there was one citation context of a base article in a first-generation article (denoted by FirstGen-article1) and while there were six citation contexts of the first-generation article in a second-generation article (denoted by SecondGen-article1). The citation pairs are shown in Figure 4.1 below.



Article	citation context one	citation context two
FirstGen-article1_SecondGen-article1	in the united states approximately 63,000 new cases of	the patient with a thyroid nodule should be asked about a fa
	in the united states approximately 63,000 new cases of	the recent 2015 american thyroid association management
	in the united states approximately 63,000 new cases of	recommendation 9 in the 2015 american thyroid association
	in the united states approximately 63,000 new cases of	the 2015 american thyroid association guidelines recommen
	in the united states approximately 63,000 new cases of	the appropriate level of thyrotrophine stimulating hormone
	in the united states approximately 63,000 new cases of	therefore women with an excellent response to therapy as d
	in the united states approximately 63,000 new cases of	although fine-needle aspiration of subcentimeter nodules in

Figure 3.1: Citation context pairs from first-generation and second-generation articles

3.4 Citation Context Similarity Classification based on Experts' Annotation

There was a need to manually classify the citation context pairs to three classes of semantic similarity (similar, somewhat similar and not similar) since there is no existing human annotated citation context semantic similarity corpora. Computer algorithm allocates semantic similarity scores between zero and one. The semantic similarity scores can be interpreted as the closer the semantic similarity score is to one, the more similar are the two texts and the closer the semantic similarity score to zero, the less similar the two texts. However, there are no known boundaries in the semantic similarity scores (between zero and one) that can be used to demarcate the three semantic similarity classes. The expert human/manual annotation objective was meant to help identify boundaries between the three semantic similarity classes so that semantic similarity scores of citation context pairs could be used to classify the citation context pairs.

Out of the 9795 pairs of citation contexts for the direct citation weighting aspect of this thesis, systematic random sampling was used to select every tenth citation context pair. 981 citation context pairs were sampled and given to two early-career biomedical experts for annotation. The first expert had completed a bachelor’s degree in three biomedical disciplines: Biochemistry, Biomedical Sciences, and Nursing. At the time of data collection, the expert was enrolled in a master’s degree program in Nursing in Western University, Canada. The second expert had completed a four-year bachelor’s degree in biomedical sciences and additional years of training in the clinical sciences to become a Physiotherapist. The second expert was also enrolled in a master’s degree in Neurological Physiotherapy at the University of Ibadan, Nigeria at the time of data annotation. The two experts were categorized as early-career experts because they had both worked for less than five years and were enrolled in master’s degree programs in medical and nursing programs.

The experts were trained on how to code the sample. During the training, they were shown the modalities of classifying citation contexts into semantic similarity classes. One of the overarching instructions was to consider the similarity of concepts/keywords in the two citation contexts as a basis for drawing semantic similarities. The annotation was a classification task into three categories of similarity: “not similar”, “somewhat similar” and “similar” based on the experts’ knowledge. Table 3.1 contains a detailed description of the three categories similarities. Therefore, a pair of citation contexts was classified as “not similar”, “somewhat similar” or “similar” based on the similarity in the concepts/keywords between the two citation contexts. Boundaries between the three classes of citation context semantic similarity were not specified, and each of the two experts decided the citation contexts' classifications.

The two experts were adequately briefed on the annotation objectives and about the citation contexts datasets. Afterwards, training was conducted online for the two experts on the semantic similarity tasks. The experts grouped citation contexts under the “not similar” class if the citation context pairs did not share similar concepts or keywords or the meanings of most or all the concepts or keywords in the citation contexts were different, thereby making the meaning of the citation context pairs not similar. On the other hand, citation context pairs were grouped under the “somewhat similar” class if the two citation contexts share some similar concepts or the keywords/concepts in the two citation contexts share some meanings. Lastly, experts classified citation contexts as “similar” if all the concepts/keywords are similar and, therefore, the citation context pair are similar in meaning. The two experts independently annotated all the sampled 981 citation context pairs.

Table 3.1: Details of human annotation's semantic similarity classes

Classes	Description
Not similar	**None of the** concepts/keywords that are identifiable in these two citation contexts are similar. Therefore, I think the two citation contexts are not similar in meaning, or they are not semantically similar.
Somewhat similar	**Some** of the concepts/keywords that are identifiable in these two citation contexts are similar. Therefore, I think the two citation contexts are somewhat similar in meaning, or they are somewhat semantically similar.
Similar	**All or most of** the concepts/keywords that are identifiable in these two citation contexts are similar. Therefore, I think the two citation contexts are similar in meaning, or they are semantically similar.

3.5 Citation Contexts Similarity Algorithm

The proposed thesis is based on citation context similarity, and this section of the methodology provides information on the computer program for the calculation of the citation context similarity score. Manually collected and pre-processed citation contexts were the inputs at this stage of the study. This stage of the study was automated using a computer program to calculate the semantic similarity score between citation contexts based on their cosine similarity measures. A python program that implements the BioSentVec sentence embeddings models was used for automating the semantic similarity measurement (Chen et al., 2019). BioSent2Vec model is the state-of-the-art for biomedical scientific publications language representation, and the model was trained on a PubMed dataset with over 28 million biomedical scientific publications. BioSent2Vec is based on the FastText's Sent2vec language text representation model which was proposed by Bojanowski, Grave, Joulin, and Mikolov (2017).

The BioSentVec converts texts to numerical representation on a multi-dimensional vector space so that the distance on the vector space represents the semantic differences between the words. It, therefore, 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. Mathematically, the cosine value is given as:

$$\cos(\theta) \frac{A.B}{|A|.|B|}$$

BiosentVec is an excellent fit for the tasks of semantic similarity because of its high accuracy in finding semantic similarities between sentence pairs when compared to human annotators. The performance of BioSentVec in finding semantic similarities was tested on the humanly annotated sentence pairs from BIOSSES (Soğancıoğlu et al., 2017) and MedSTS (Wang et al., 2018).

Thresholds for the three classes of semantic similarities of citation contexts (“similar”, “somewhat similar”, and “not similar”) were obtained using the citation contexts semantic similarity classification by human experts that was described in the previous section. The weights of the citation contexts in each of the semantic similarity classes from human expert annotation were obtained, using histogram of the distribution of the three classes and through iteration, the boundaries of the three classes were obtained.

3.6 Existing Citation Metrics and Citation sentiment

Included in this thesis is a comparison between earlier citation metrics with the proposed semantic similarity citation context weighting method. All the selected existing metrics - number of citations received by a publication, number of citation mentions and number of multiple citation mentions, and sum of multiple citation mentions- were derivatives of citation and citation mentions, except number of positive citation sentiment. Citation sentiments were obtained using a python implementation of pattern.en. Pattern.en was created by Smedt & Daelemans, (2012), researchers at the Centre for Computational Linguistics and Pyscholinuistics (CLIPS), University of Antwep, Belgium as python package for natural language processing research in both scientific and non-scientific settings.

3.7 Direct Citation Data Weighting

Weights were allocated to citation context similarity based on thresholds that were established using the experts' annotation. Two unique, not similar, citation contexts were allocated a weight of two. Therefore, each of the two unique citation contexts was assigned a weight of one. The theoretical consideration for this is the simple counting of unique items in mundane activities or real world, where one is added whenever a unique item is added. Two perfectly similar citation contexts were allocated weight of one. In theory, the two similar citation contexts are similar to the extent that the two citation contexts could be regarded as one. Two somewhat similar citation contexts were allocated a weight of 1.5. In theory, two somewhat similar citation contexts are entitled to a weight that is less than those of unique citation contexts and greater than those of similar citation contexts. Therefore, each of the somewhat similar citation contexts was allocated a weight of 0.75. With this weighting system, one citation context did not receive a weight that is greater than one.

This is a simple way of allocating weights for the two mentions of a cited paper in a citing paper. However, if there are more than two mentions, the number of citation contexts will be greater than two. Hence, in the next paragraph, we present a general algorithm for allocating weights to a cited paper that is mentioned n times in the citing paper.

Let us start with a cited paper that is mentioned twice in the citing paper. The first citation context is allocated a weight of one, but the additional weight allocated to the second citation context is determined by its semantic similarity to the first. Therefore, a weight of 0 is allocated if the second citation context was similar to the first, weight of 0.5 is allocated it the two citation contexts are somewhat similar, and a weight of 1 was allocated if the two citation contexts are not similar. For a cited paper that is mentioned three times (*n*=3) M_a, $M_{b,}$ M_c, the citation context pairs are $M_a|M_b$, $M_a|M_c$ and $M_b|M_c$. For the last citation context M_c, it could only be allocated the weight of one. However, its weight depends on its comparison with citation contexts M_a and M_b. In theory, the maximum weight that could be obtained for M_c wrt M_a =1 and M_c wrt M_a=1, adding the two maximum weights and dividing by 2 ((1+1)/2)=1. So, the weight of a citation with three mentions is calculated as

$$= \text{weight of Ma} + \text{weight of Mb|Ma} + \left(\frac{weight\ of M_c \mid M_a + \text{weight of } M_c \mid M_b}{2}\right)$$

For a citation that is mentioned four times M$_a$, M$_b$, M$_c$, M$_d$ the citation context pairs are M$_a$|M$_b$, M$_a$|M$_c$, M$_b$|M$_c$, M$_a$|M$_d$, M$_b$|M$_d$ and M$_c$|M$_d$. and the weight was calculated as

$$\text{weight of Ma} + \text{weight of Mb} \mid \text{Ma} + (\frac{\mathit{weight\ of}\, M_c \mid M_a + \text{weight of } M_c \mid M_b}{2})$$

$$+ (\frac{\mathit{weight\ of}\, M_d \mid M_a + \text{weight of } M_d \mid M_b + \text{weight of } M_d \mid M_c}{3})$$

For a citation that is mentioned five times M$_a$, M$_b$, M$_c$, M$_d$, M$_e$ the citation context pairs are M$_a$|M$_b$, M$_a$|M$_c$, M$_b$|M$_c$, M$_a$|M$_d$, M$_b$|M$_d$, M$_c$|M$_d$, M$_a$|M$_e$, M$_b$|M$_e$, M$_c$|M$_e$, and M$_d$|M$_e$, and the weight was obtained as

$$\text{weight of Ma} + \text{weight of Mb}|\text{Ma} + (\frac{\mathit{weight\ of}\, M_c \mid M_a + \text{weight of } M_c \mid M_b}{2})$$

$$+ (\frac{\mathit{weight\ of}\, M_d \mid M_a + \text{weight of } M_d \mid M_b + \text{weight of } M_d \mid M_c}{3})$$

$$+ (\frac{\mathit{weight\ of}\, M_e \mid M_a + \text{weight of } M_e \mid M_b + \text{weight of } M_e \mid M_c + \ \mathit{weight\ of}\, M_e \mid M_d}{4})$$

In general, for a citation that is mentioned *n* times M$_1$, M$_2$, M$_3$, M$_4$,… M$_{n-1}$, M$_n$ the citation context pairs are M$_1$|M$_2$, M$_1$|M$_3$, M$_1$|M$_4$, … M$_{n-1}$|M$_n$, and the weight was obtained as

$$\text{weight of } M_1 + \text{weight of } M_2 \mid M_1 + \left(\frac{\mathit{weight\ of}\, M_3 \mid M_1 + \text{weight of } M_3 \mid M_2}{2}\right)$$

$$+ \left(\frac{\mathit{weight\ of}\, M_4\ \mathit{wrt}\ M_1 + \text{weight of } M_4 \mid M_2 + \text{weight of } M_4 \mid M_3}{3}\right) + \cdots$$

$$+ (\frac{\mathit{weight\ of}\, M_n \mid M_1 + \text{weight of } M_n \mid M_2 + \cdots + \ \mathit{weight\ of}\, M_n \mid M_{n-1}}{n-1}$$

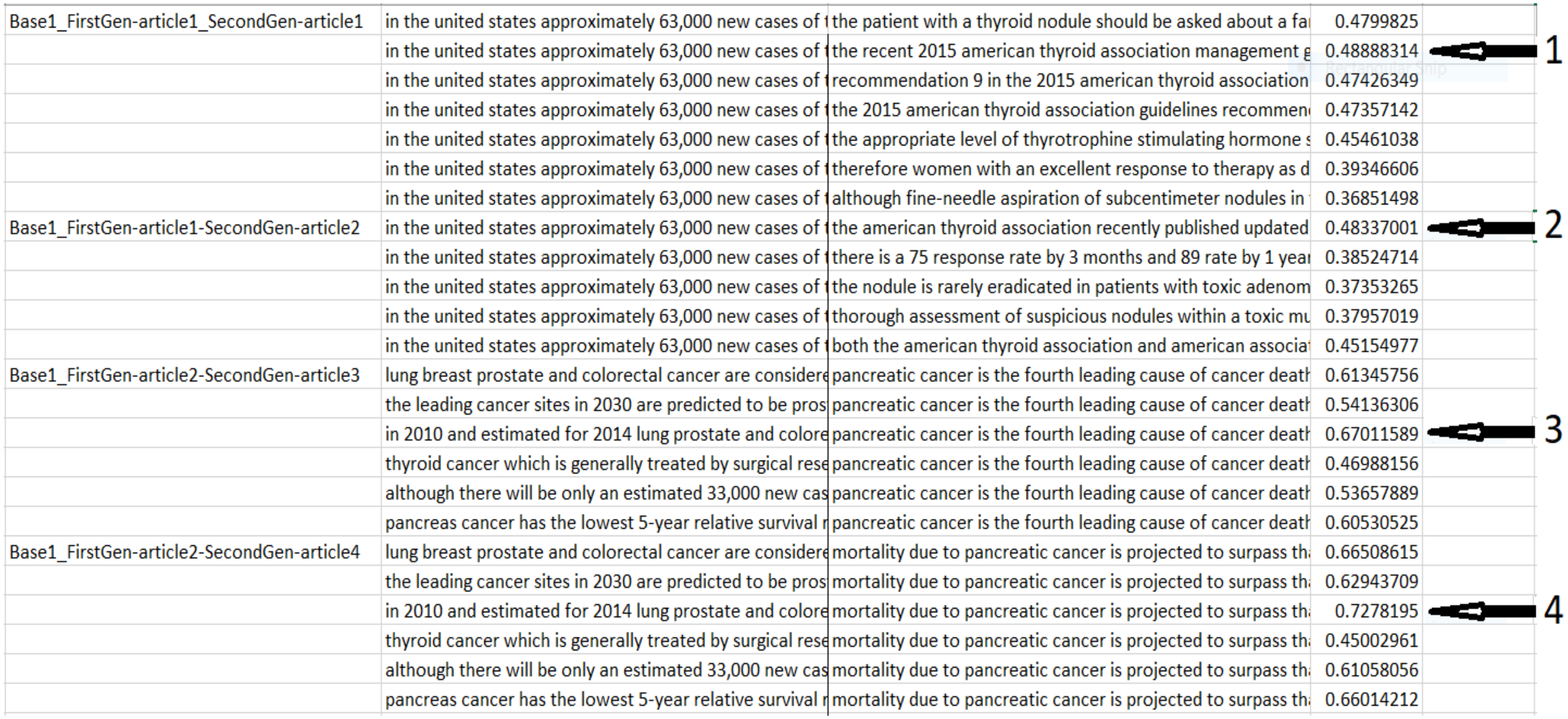
3.8 Indirect Citation Data Weighting

Firstly, in order to determine the amount of residual citation that should accrue to a paper article from a second generation, we examined the number of mentions of the base article in the first generation citation and the number of mentions of the first generation citations in the second generation citation. Secondly, in order to determine the amount of residual citation that should accrue to a base article from a third generation citation, we examined the number of mentions of the base article in the first generation citation and the number of mentions of the second generation paper in the third generation citations. Thirdly, in order to determine the amount of residual citation that should accrue to a base article from a fourth generation citation, we examined the number of mentions of the base article in the first generation citation and the number of mentions of the third generation paper in the fourth generation citations. Finally, in order to determine the amount of residual citation that should accrue to a base article from a fifth generation citation, we examined the number of mentions of the base article in the first generation citation and the number of mentions of the fourth generation paper in the fifth generation citations.

In each of the four cases above, citation contexts of the two sets of mentions were collected and were paired up. For example, let the base article be mentioned in a first generation citation three times, and the first generation citation be mentioned in the second generation citation four times,. Let us represent these mentions as follows: M_{a_01}, M_{b_01},and M_{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 3.2, 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.



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 year	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 th	0.66508615	
	the leading cancer sites in 2030 are predicted to be pros	mortality due to pancreatic cancer is projected to surpass th	0.62943709	
	in 2010 and estimated for 2014 lung prostate and colore	mortality due to pancreatic cancer is projected to surpass th	0.7278195	4
	thyroid cancer which is generally treated by surgical rese	mortality due to pancreatic cancer is projected to surpass th	0.45002961	
	although there will be only an estimated 33,000 new cas	mortality due to pancreatic cancer is projected to surpass th	0.61058056	
	pancreas cancer has the lowest 5-year relative survival r	mortality due to pancreatic cancer is projected to surpass th	0.66014212	

Figure 3.2: Citation context pairs for four articles with marked highest semantic similarity measures

3.9 Statistical Analysis

We used tables and charts to illustrate the data collected and appropriate measures of central of tendency were also used to describe the data. In testing the hypothesis, non-parametric tests were preferred over parametric tests due to the features of the datasets. Specifically, Spearman’s Rank Correlation and Kruskal-Wallis tests were used. The level of significance was set at 0.05.

Chapter 4

4 Results

The chapter presents results from the analysis of the datasets using the methods that were described in the previous chapter. The results are presented in four sections. Results of the human expert annotation are presented in the first section. Descriptive Statistics and metadata of the direct citation datasets are presented in the second section of this chapter. Result of the comparison between the proposed semantic similarity-based citation weights and existing metrics are presented in the third section. Lastly, results of the analysis of the indirect citation datasets are presented in section 4 of this chapter.

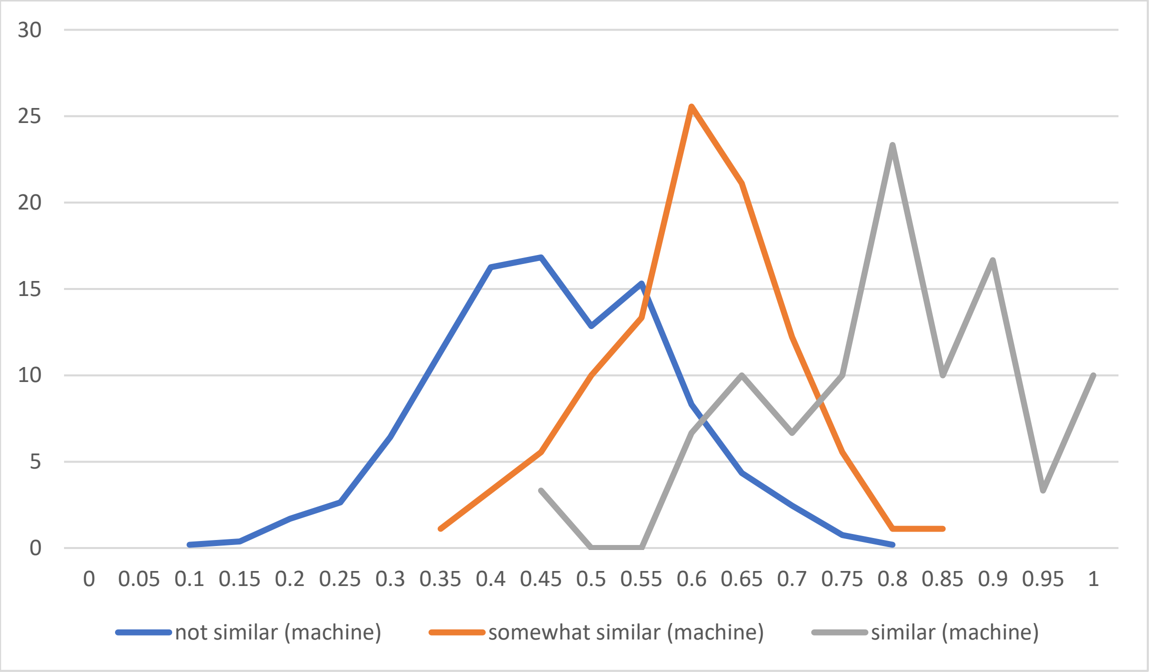
4.1 Human Annotation and Class Thresholds

The two experts independently annotated all the sampled 981 citation context pairs, and the inter-coder agreement of the two coders was 66.16% (649/981) based on the percentage agreement and 0.27 Cohen Kappa score. Semantic similarity scores were obtained by computer algorithm for citation context pairs that were classified into three classes by human annotators. Identifying the boundaries for three classes for semantic similarity score obtained from computer algorithm was done manually using a histogram (see Figure 4.1) and a frequency table (see Table 4.1). Data points between two classes (somewhat similar/similar, and not similar/somewhat similar) that returned the highest percentages of true positives for the two classes whose boundaries were chosen as the boundary between the two classes.

Table 4.1: Distribution of the semantic similarity scores vs Human Classification

interval	*not similar (machine)*	*somewhat similar (machine)*	*similar (machine)*
0.1	1		
0.15	2		
0.2	9		
0.25	14		
0.3	34		
0.35	60	1	
0.4	86	3	
0.45	89	5	1
0.5	68	9	0
0.55	81	12	0
0.6	44	23	2
0.65	23	19	3
0.7	13	11	2
0.75	4	5	3
0.8	1	1	7
0.85		1	3
0.9			5
0.95			1
1			3

Figure 4.1: Histogram of *S*emantic Similarity Scores *versus H*uman *C*lassification



The boundaries for the citation context classifications are displayed in Table 4.2. For the "not similar" semantic class, the boundary was pegged at $0<x<0.51$. At this boundary, 71.46% of the citation context pairs that were classified as "not similar" by humans were classified correctly. If the threshold increased, more citation context pairs in the "not similar" will be captured but simultaneously, more citation context pairs of the "somewhat similar" class will also be wrongly classified, thereby increasing the recall and reducing the precision. This process was repeated for the two other classes.

Table 4.2: Boundaries between the three classes of semantic similarity

Classes	Not similar	Somewhat similar	similar
Boundary for computer algorithm	<0.51	>=0.51 and <0.71	>=0.71
Percentage accuracy for computer algorithm	71.46%	72.22%	73.33%

4.2 Direct Citation Data Description

The 100 sampled articles received 7317 citations (max.=179, min.=31, average=73.17); that is, citation contexts were extracted from 7317 citing articles. Please, note that citations received refers to the number of citing articles that were available during the data collection. The 100 sampled articles originally received 8208 citations but 891 articles were removed either because they were not available during data collection, the cited articles were not found in the full text, or the cited articles in-text citations were part of Tables or Figures. Details about the 100 articles can be found in Appendix A.

The frequency distribution of mentions of the 100 cited articles as in-text citations in the 7317 citing articles are displayed in Table 4.3. Most of the citations of the publications (73.02%) were mentioned once in the citing articles, which means that only 26.98% of the citations were included in the citation context similarity weighting. There were 11,234 citation contexts or mentions, and the highest number of mentions was 31.

Table 4.3: Number of Citation mentions

Number of citation mentions	Frequency	Percentage
1	5333	72.89
2	1174	16.045
3	399	5.45
4	187	2.56
5	77	1.05
6	51	0.70
7	30	0.41
8	16	0.22
9	11	0.15
10	14	0.19
11	6	0.082
12	8	0.109
13	2	0.027
14	2	0.027
15	1	0.014
16	2	0.027
19	2	0.027
25	1	0.014
31	1	0.014
Total	7317	

4.2.1 Results of the Existing Citation weights

The distributions of the number of citations and citation mentions by the 100 cited articles are presented in Figure 4.2 and Figure 4.3, respectively, which show that the distributions are approximately normal. The normal distributions of the samples justify the use of mean as the measure of central tendency. The distributions of the sum of multiple citation mentions, the number of multiple citation mentions and the number of positive citation sentiments in Figure 4.4, Figure 4.5, and Figure 4.6, respectively, show that the three datasets are asymmetrical.

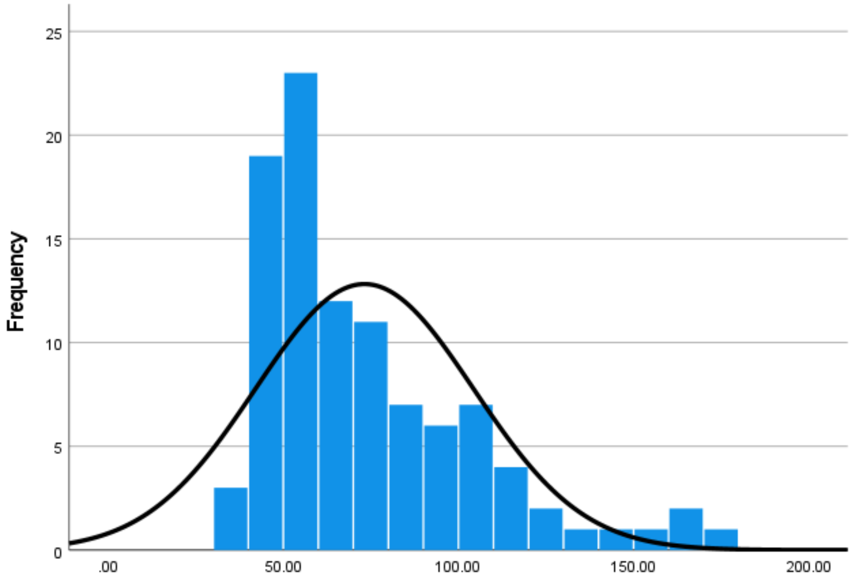


Figure 4.2: Distribution of the citation numbers received

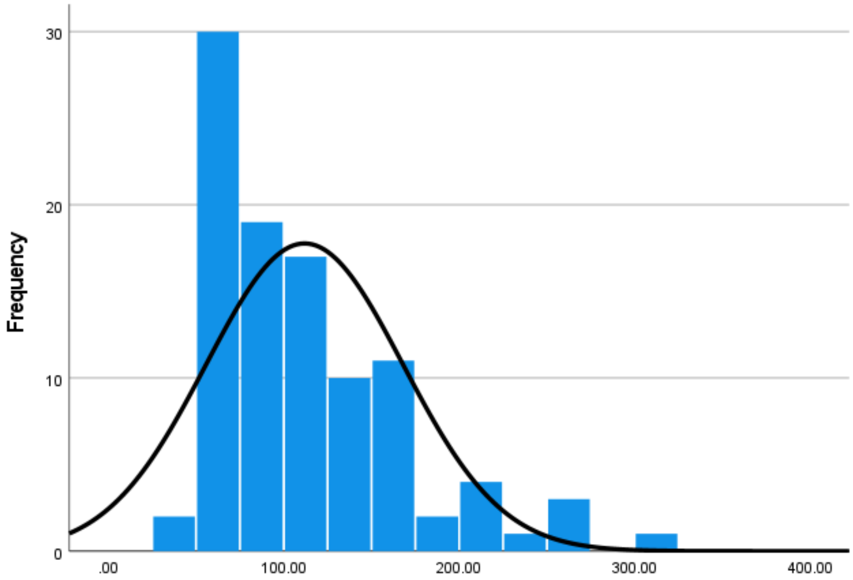


Figure 4.3: Distribution of citation mentions per cited article

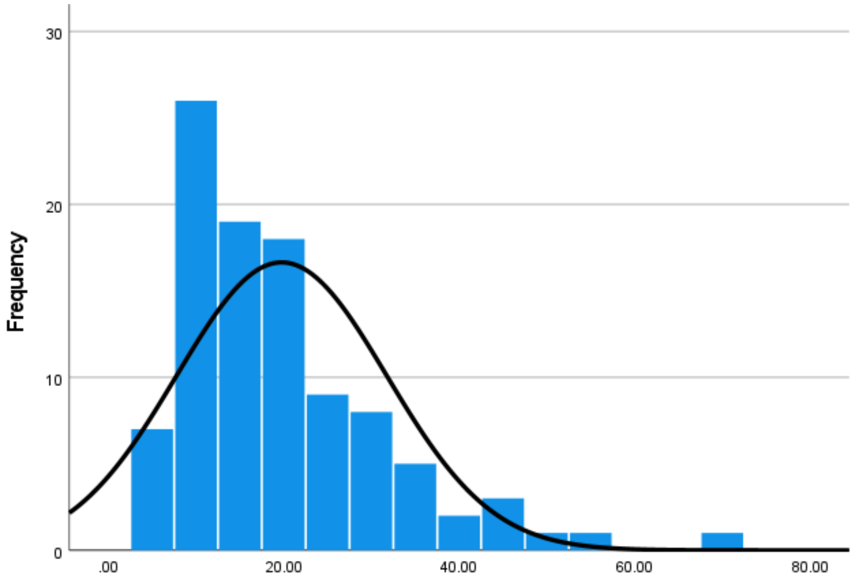


Figure 4.4: Distribution of the number of multiple citation mentions

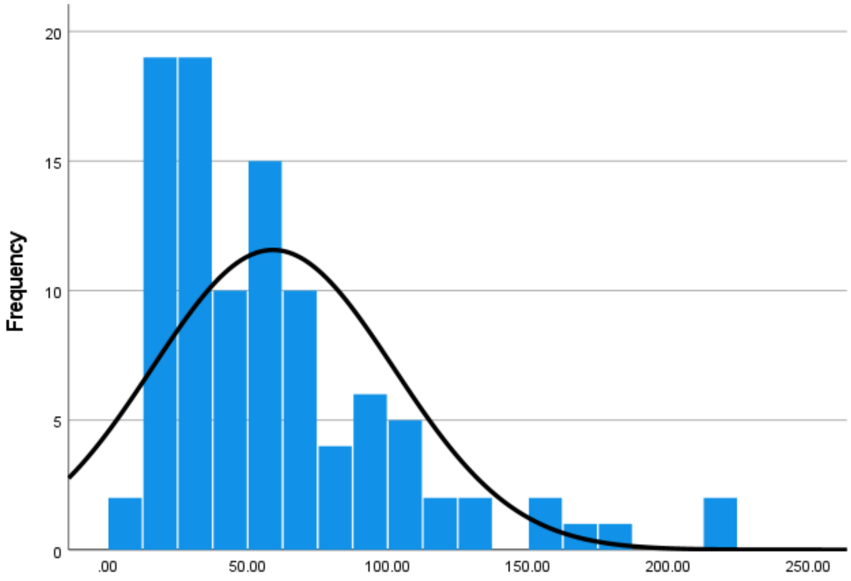


Figure 4.5: Distribution of the sum of multiple citation mentions

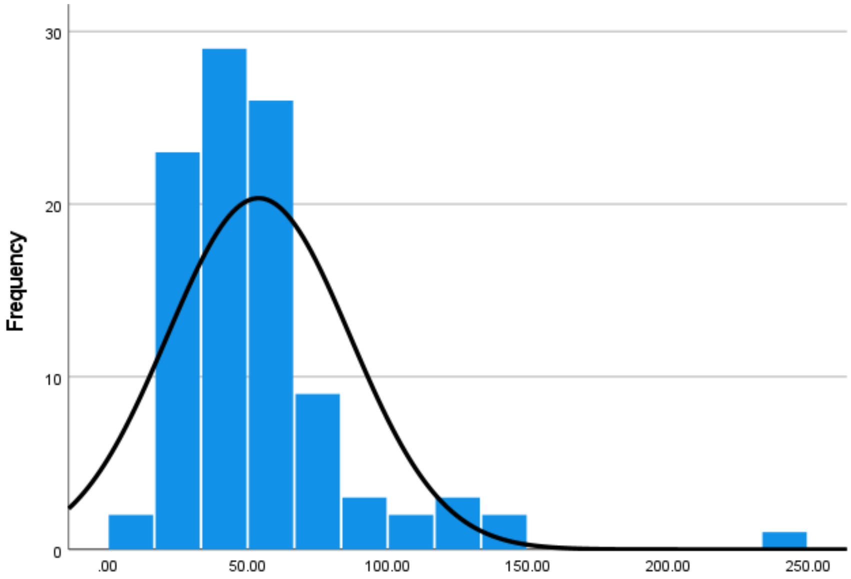


Figure 4.6: Distribution of the number of positive citation context sentiment per cited article

4.2.2 Relationship between the proposed semantic similarity-based citation weight and existing metrics?

Results in this sub-sections are tied to Research Question 1 guiding this thesis. The distribution of semantic similarity-based citation weights is presented in Figure 4.7. The figure shows that the distribution is approximately normal. Thus, mean was presented as the measure of central tendency. The sampled article received a mean of 99.72 (maximum=248.96, minimum=40.29) semantic similarity-based citation weights.

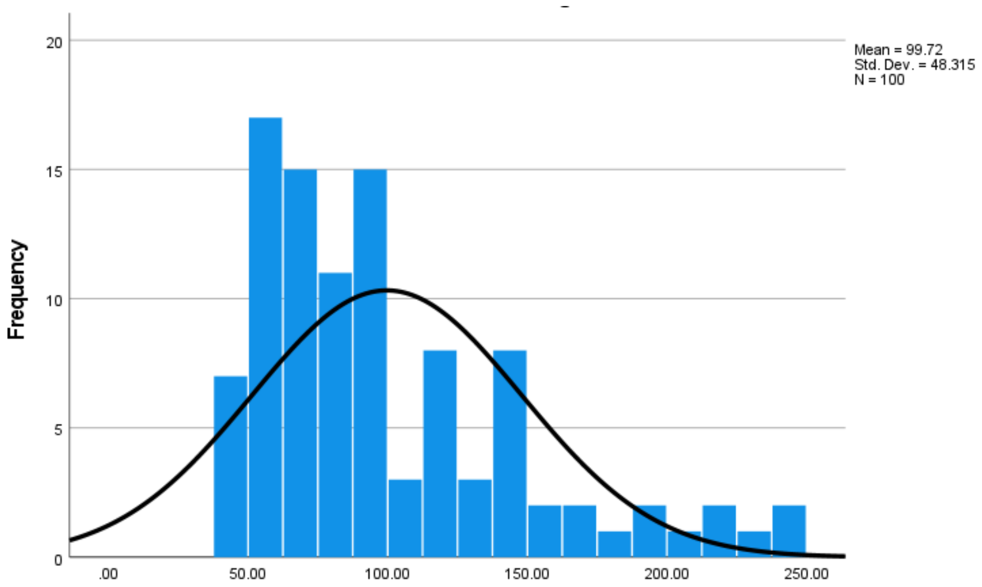


Figure 4.7: Distribution of the proposed semantic similarity-based citation weights

In theory, the proposed semantic similarity-based citation context weight is greater than or equal to the citation count and less than or equal to the number of citation mentions. A comparison of the proposed semantic-similarity-based citation context weights, citation count and the number of citation mentions of all the 100 sampled articles is visualized in Figure 4.8 to give a holistic picture of the validity of the proposed citation weight in theory. Figure 4.8 shows that the implementation of the proposed semantic similarity-based citation context weight from this doctoral thesis produced a valid result. The sampled articles received semantic similarity-based weights that are less than the number of citation mentions but greater than the number of citations. This shows that the semantic similarity-based citation weight implementation is accurate since the weights discounted ordinary citation mention count. However, the proposed semantic similarity-based citation weight in this thesis places a premium on the citation context's uniqueness. Therefore, for every multiple mentioned citation, there is at least a unique citation mention and a maximum number of unique citation mentions equal to the number of citation mentions.

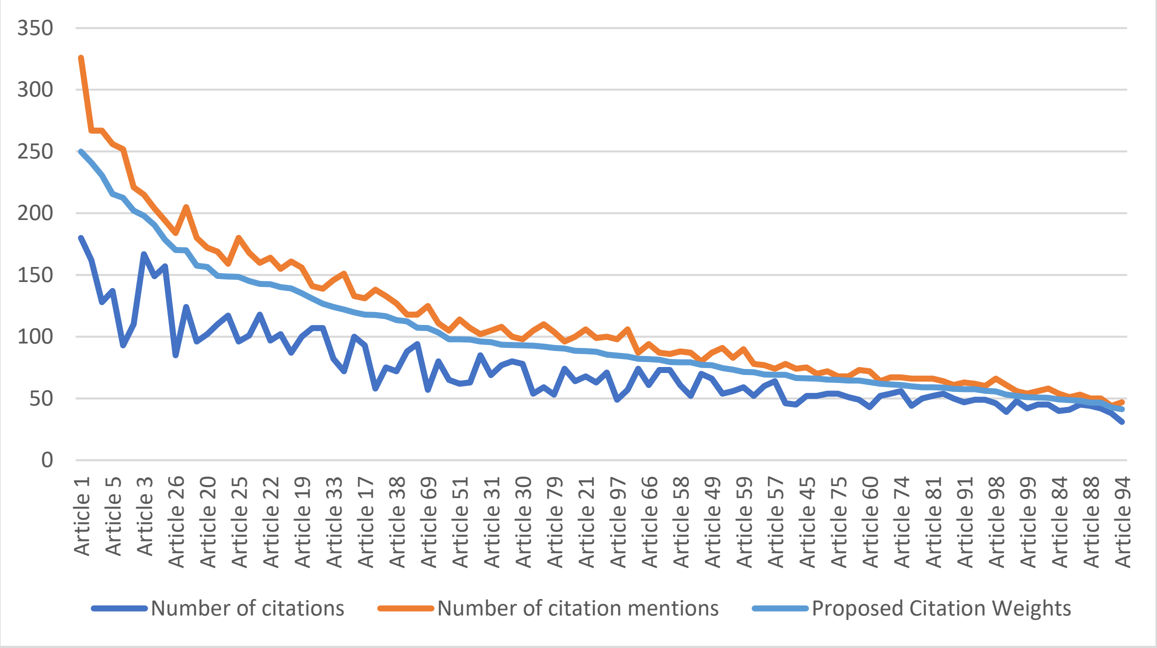


Figure 4.8: Sampled articles ranked by citation number

The following hypotheses were tested using Spearman’s rho correlation statistical test. The results of the tests are displayed in Table 4.4.

Hypothesis 1$_0$: There is no correlation between the number of citations and the proposed citation context similarity-based citation weight

Hypothesis 2$_0$: There is no correlation between the number of citation mentions and the proposed citation context similarity-based citation weight

Hypothesis 3$_0$: There is no correlation between the number of multiple citation mentions and the proposed citation context similarity-based citation weight

Hypothesis 4_0: There is no correlation between the sum of multiple citation mentions and the proposed citation context similarity-based citation weight

Hypothesis 5_0: There is no correlation between the number of positive sentiments and the proposed citation context similarity-based citation weight

Table 4.4: Correlations between the proposed citation weighting methods and existing metrics

	Metric	Correlation co-efficient	Significant level (2 tailed)	Statistically significant
Proposed Metric	Number of citations	.93	.00	Yes
	Number of citation mentions	.99	.00	Yes
	Number of multiple citation mentions	.89	.00	Yes
	Sum of multiple citation mentions	.89	.00	Yes
	Number of positive sentiments	.86	.00	Yes
	Proposed Metric with semantic similarity scores (without citation count)	.83	.00	Yes

The results of Spearman’s rho correlation tests in Table 4.4 show there is a strong, positive, and significant relationship between the proposed measure and each of the number of citations received, the number of citation mentions, the number of multiple citation mentions, the sum of multiple citation mentions, and the number of positive citation sentiments. Therefore, **Hypothesis 1_0**, **Hypothesis 2_0**, **Hypothesis 3_0**, **Hypothesis 4_0**, and **Hypothesis 5_0** are rejected.

4.3 Residual Citation

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 10 base articles can be found in Table 4.5. Included in Table 4.5 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 4.5: Indirect Citations Statistics

		times	sample size per generation					citation contexts no/generation				
	Table	cited	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 4.6: 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 4.7. below. First, the distributions of the semantic similarity measures on histogram graphs were inspected visually for normality. The distributions are presented in Figure 4.9, Figure 4.10, Figure 4.11 and Figure 4.12, and they are symmetrical in shape. Table 4.14 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 generations.

Table 4.7: Averages of the residual citations received from the second to the fifth citation generations

	Second Generation	Third Generation	Fourth Generation	Fifth Generation
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

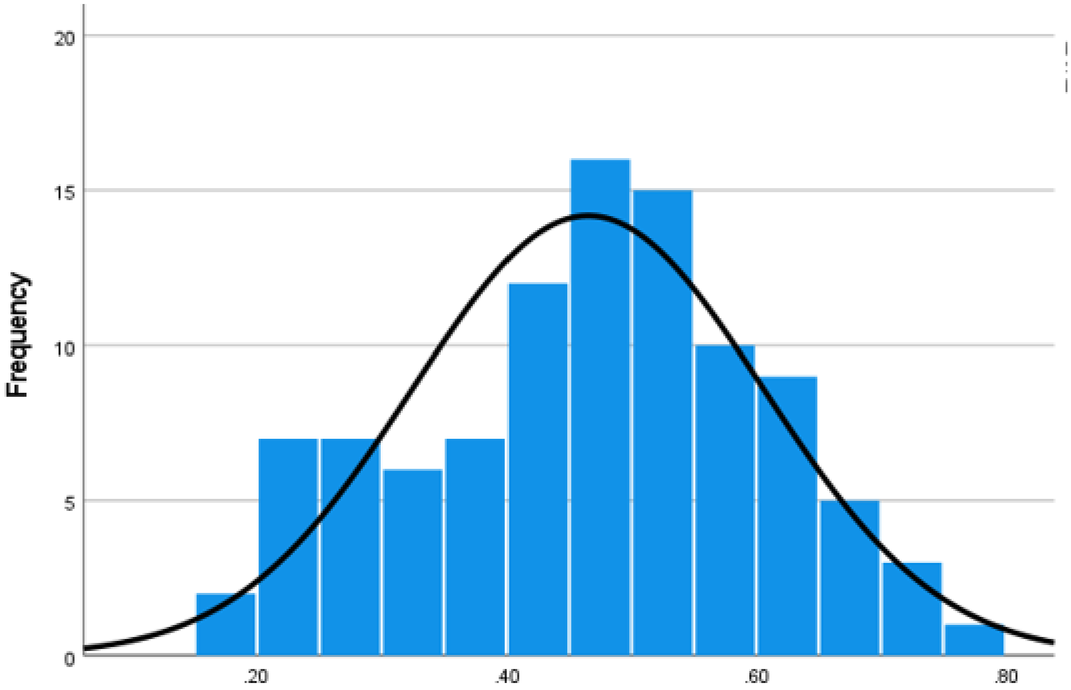


Figure 4.9: Distribution of the residual citations received from the second generation citations

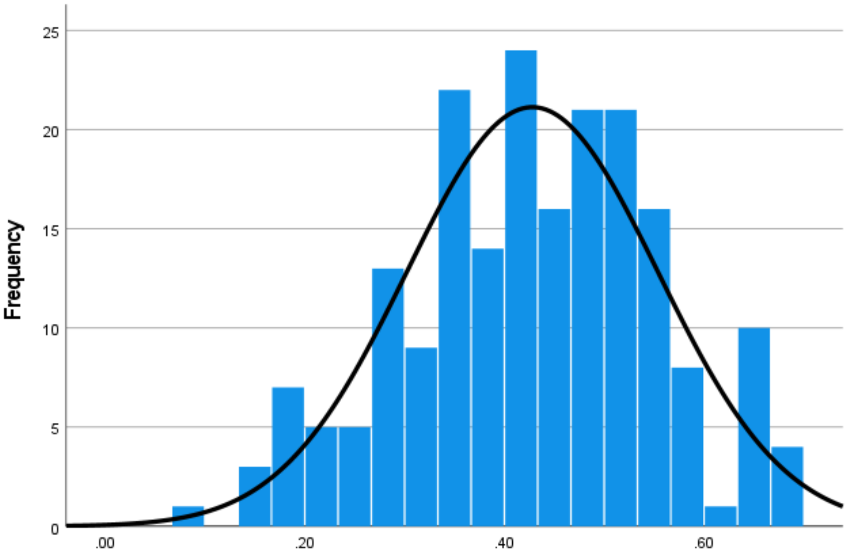


Figure 4.10: Distribution of the residual citations received from the third generation citations

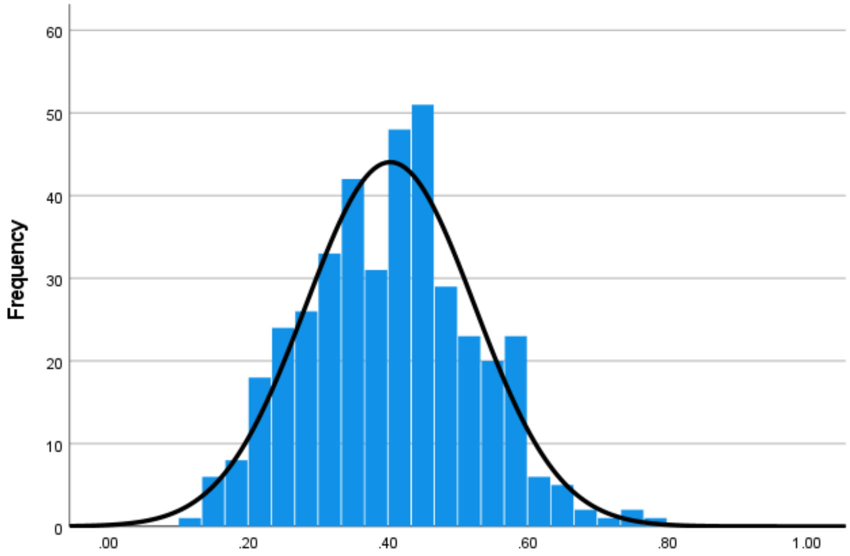


Figure 4.11: Distribution of the residual citations received from the fourth generation citations

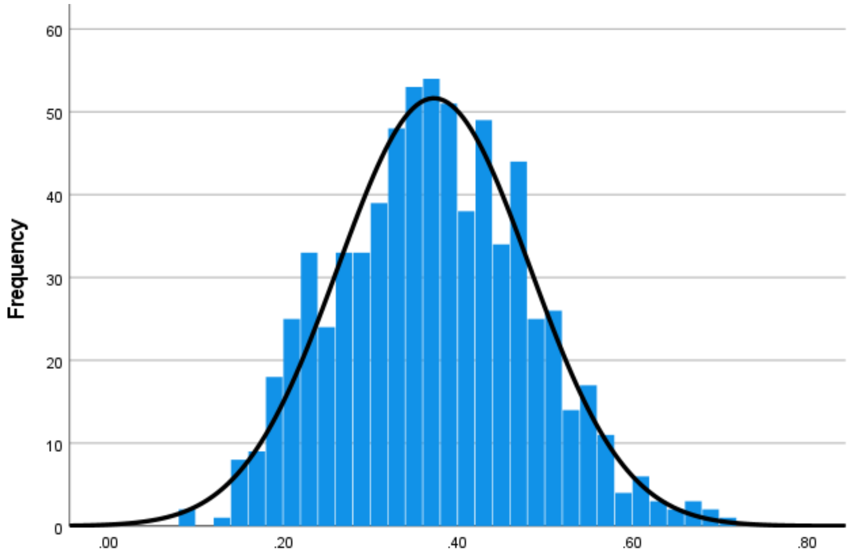


Figure 4.12: Distribution of the residual citations received from the fifth generation citations

4.3.1 Residual Citation Patterns between the cited documents and their nth Generation Citations

This subsection presents the results on the Research Question 4. The semantic similarity-based residual citation weights were categorized using the thresholds that were specified in Section 4.1 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 a weight of one.

Categorization of the weights (see Table 4.8) 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 4.8: 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 4.9. 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 generation except the 5th 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 4.9: Non-zero indirect Citation weights

Base articles	Non-zero residual weights			
	2nd Generation	3rd Generation	4th Generation	5th 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%

4.3.2 Differences in the residual citations between the generations of citation

This sub-section contains the results on Research Question 2. From the observations in Table 4.7, 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 4.9 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 residual citations from the base article, the result of the averages of the semantic similarity measure shows that the residual citations received by the base article continuously reduced as the generations of citations increased. Therefore, **Hypothesis 6_0** was stated to guide this thesis. Inferential statistics was therefore performed to confirm if the observed differences in the averages are significant.

Since the data is continuous, a recommended statistical test is the analysis of variance (ANOVA). It was tested to determine 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 4.13, Figure 4.14, Figure 4.15 and Figure 4.16, 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 4.9. 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 4.10: 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 4.11 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 4.11: 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 4.12: 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. Result of the pairwise comparison test is displayed in Table 4.13.

Table 4.13: Pairwise Comparisons of the Semantic Similarity Score categories

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.

The result of the pairwise comparison as shown in Table 4.13, The difference between first-generation and every other generation under consideration is significant.

4.3.3 Difference between the proposed residual citation and the cascading citation weights

This sub-section contains the results on Research Question 3. 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. At the second phase, cascading citation weight per indirect citation was compared to the average semantic similarity score.

4.3.3.1 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 *4.13*: Second Generation Indirect Citation Weights Comparison**Figure 4.13. 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.

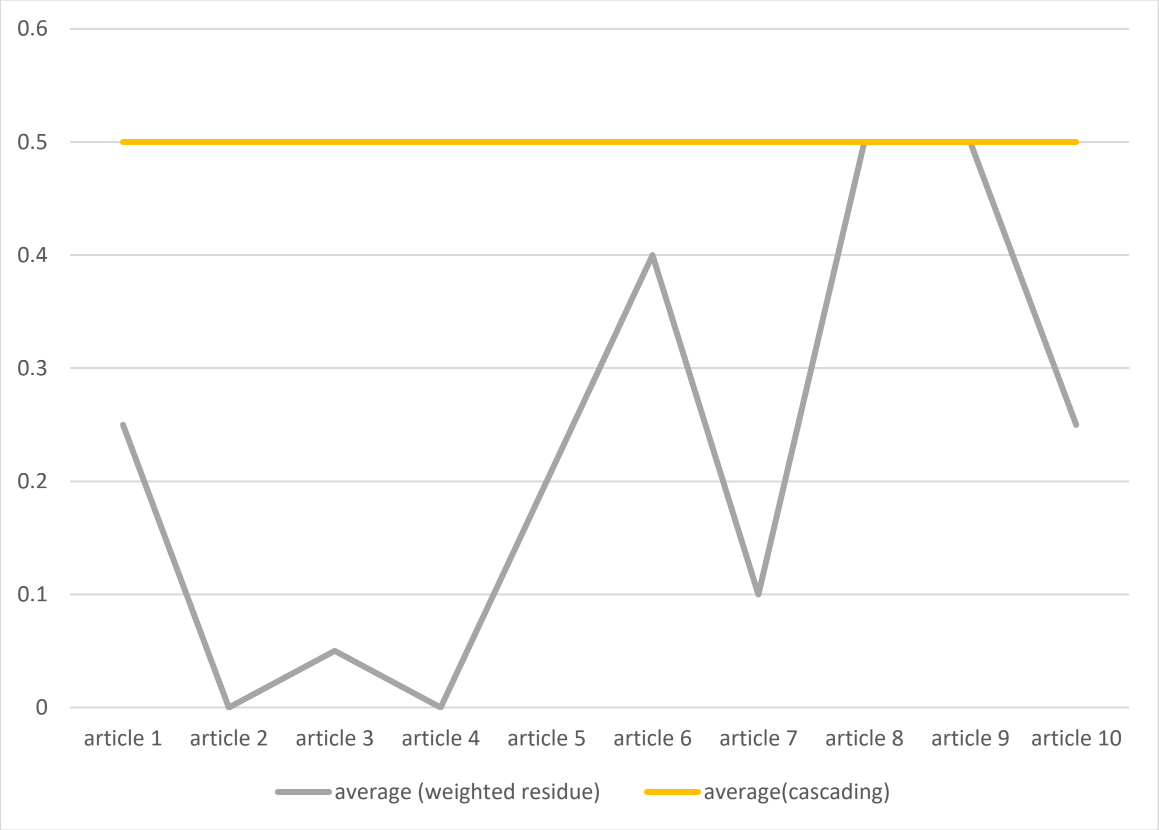


Figure 4.13: Second Generation Indirect Citation Weights Comparison

The comparison between the proposed citation weights and the cascading citation system at the third generation is visualized in Figure 4.14. 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 at 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.

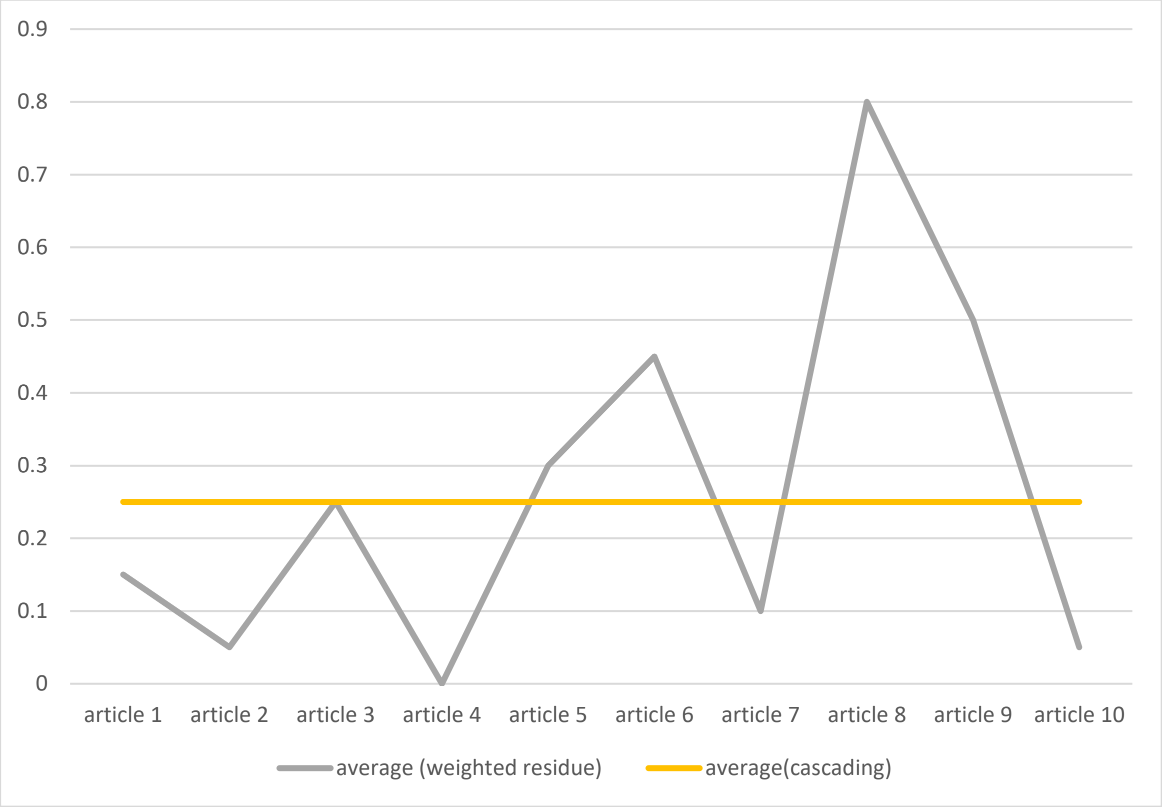


Figure 4.14: Third Generation Indirect Citation Weights Comparison

The comparison between the proposed citation weights and the cascading citation system at the fourth generation is visualized in Figure 4.14. 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 at the third generation. The average value of residual citations per base article continued to increase, though the semantic similarity score average reduced.

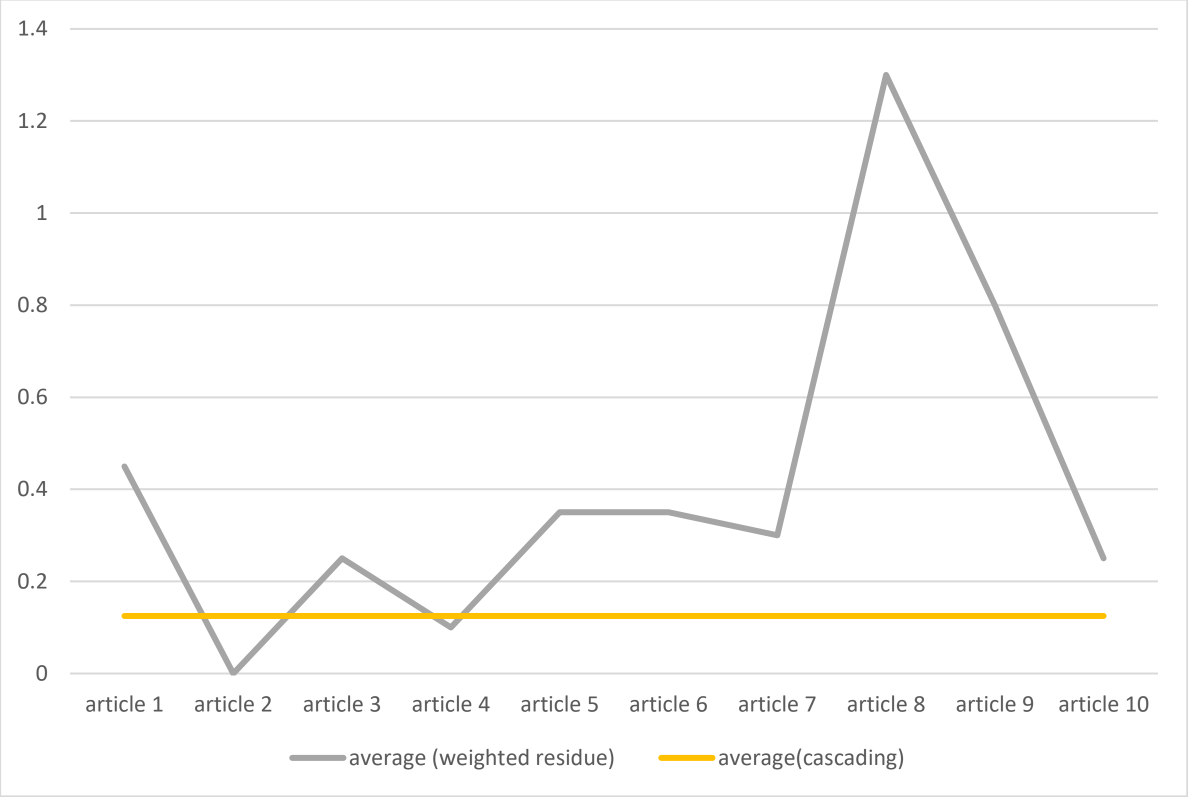


Figure 4.15: Fourth Generation Indirect Citation Weights Comparison

The comparison between the proposed citation weights and the cascading citation system at the fifth generation is visualized in Figure 4.14. 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 at the fifth generation.

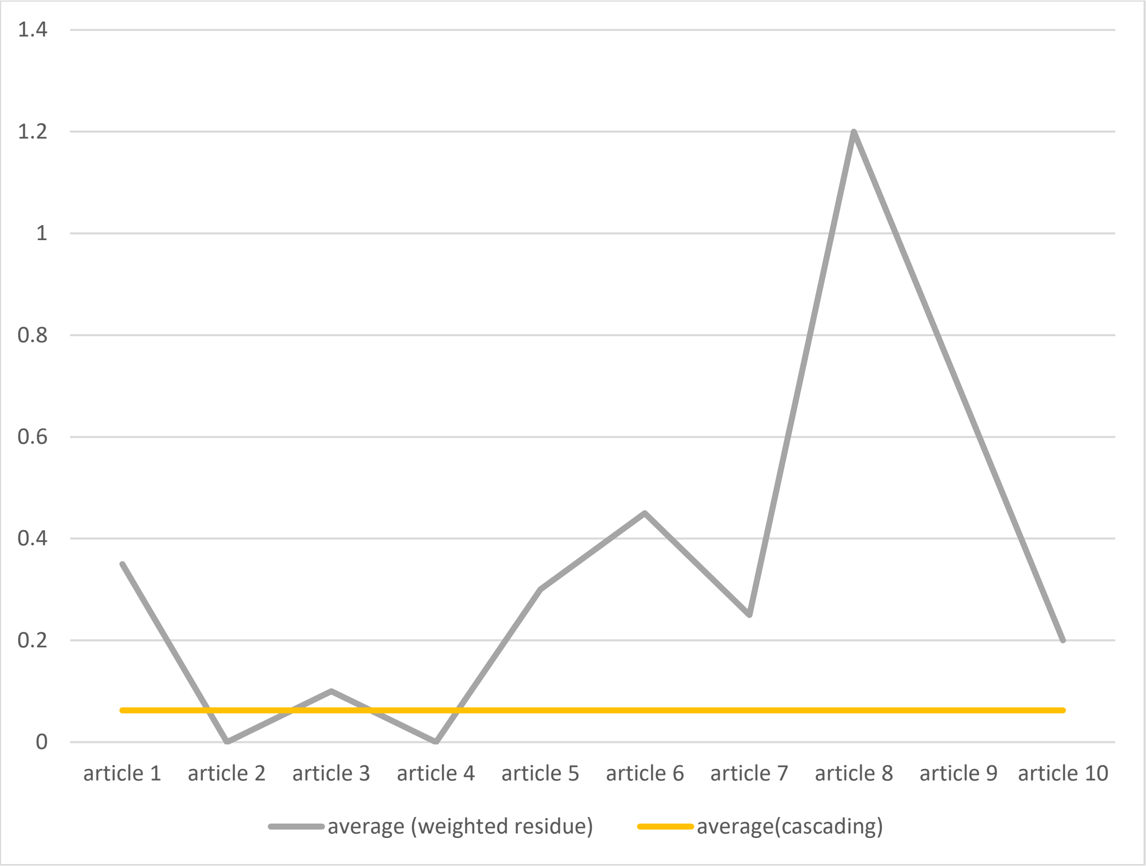


Figure 4.16: Fifth Generation Indirect Citation Weights Comparison

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

Hypothesis 7₀, Hypothesis 8₀, Hypothesis 9₀, and **Hypothesis 10₀** 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 7₀** 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 4.9, Figure 4.10, Figure 4.11 and Figure 4.12), 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 4.14. 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 4.14: One-Sample T-Test result for the Average Residual Citation Score

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Second generation	-9.61	42	.000	-.058	-.0697	-.0460
Third generation	38.95	88	.000	.14	.1321	.1461
Fourth generation	89.04	15	.000	.25	.2443	.2553
Fifth generation	138.59	2449	.000	.27829	.2743	.2822

One sample T-test result, presented in Table 4.14 shows from second to the fifth generation, the residual citations accrued by the sampled articles derived from the proposed semantic similarity method and the old cascading citation system are significantly different. The residual citations were mostly under-estimated by the cascading citation system.

Chapter 5

5 Discussion

This chapter discusses the findings as they relate to the research questions that were stated in chapter 1 of this thesis. Contributions of this study to knowledge in the relevant fields that may not be captured under the research questions were also included in the discussion section.

5.1 Research Question 1: What is the relationship between the proposed semantic similarity-based citation context weight and existing metrics?

The proposed semantic similarity-based citation weight was compared with the following existing metrics; the number of citations, number of citation mentions, number of multiple citations, the sum of multiple citation mentions and number of positive sentiments. The comparisons were based on three methods. First, the proposed weight was compared with the number of citations and the number of citation mentions using visualization in the methodology section. Secondly, the proposed metric was compared to the existing metrics using Spearman’s rho statistical test to investigate the relationship between the proposed metric and the existing. In addition, it was investigated if the relationships between the proposed and existing metrics were statistically significant. Thirdly, the rankings of the sampled articles obtained when based on the proposed metric was compared with the rankings obtained when based on each of the existing metrics. Ranking the sampled articles gave a different picture of the relationship between the sampled articles. The first twentyfive of the sampled articles' ranking by the proposed semantic similarity-based citation weight is presented in Table 5.1.

Table 5.1 shows significant rank changes among the 25 top cited publications when they were ranked using the proposed semantic similarity-based metric. For instance, article 25 that received only 93 citations moved up 20 places because of its influence in the citing publications with high number of citation mentions and multiple mentioned citations. Article 25 moved above articles that received 70 citations more. This in turn reflected in its high semantic similarity-based metric. Other notable upward rank change occurred with the ranked 28 publication that moved up by 18 positions. Notable downward rank changes occurred at the article 13 and article 14 publications that moved by ten positions, despite receiving 20 more citations than many publications above it in the new position. Article 13 and article 14 moved significantly downward because they received lower number of citation mentions and multiple citation mentions.

Table 5.1: The first 25 articles ranked by the proposed weighted citation

Rank by citation number	Diff in rank	Semantic similarity-based citation weight	citation	citation mentions	Number mentions>1	Sum mentions>1	Positive sentiment citation
rank 1	0	249.96	180	326	70	216	241
rank 3	1	241.03	162	267	51	156	135
rank 7	4	230.58	128	267	47	186	118
rank 6	2	215.72	137	256	47	166	119
rank 25	20	212.52	93	252	55	214	127
rank 11	5	202.24	110	221	41	152	113
rank 2	5	197.92	167	215	29	77	146
rank 5	3	190.58	149	204	34	89	114
rank 4	5	178.5	157	194	29	66	39
rank 28	18	170.14	85	184	32	131	94
rank 8	3	169.88	124	205	41	122	68
rank 21	9	157.67	96	180	32	116	56
rank 16	3	156.56	102	172	32	102	54
rank 12	2	149.21	110	169	33	92	64
rank 10	5	148.75	117	159	20	62	90
rank 22	6	148.36	96	180	45	129	53
rank 17	0	145.02	101	168	33	100	78
rank 9	9	142.73	118	160	26	68	79
rank 20	1	142.48	97	164	32	99	73
rank 15	5	140.26	102	155	23	76	58
rank 27	6	139.23	87	161	35	109	80
rank 19	3	135.33	100	156	33	89	59
rank 13	10	130.91	107	141	21	55	80
rank 14	10	126.74	107	139	21	53	57
rank 30	5	124.04	82	146	31	95	53

This study showed a strong, positive and significant relationship between the proposed citation weight and the number of citations received by the sampled articles. This observation implies a linear relationship between the number of citations and the proposed semantic similarity-based citation weights, and the proposed citation weight is an alternative to the number of citations. The number of citations is an easy to compute metric, though superficial. On the other hand, the proposed metric is more complex to compute but more nuanced as it depicts a degree of contribution because it is a measure that is based on the analysis of the contribution of the cited article in the citing article. Secondly, the proposed metric serves the practical use of citation, given that citations are supposedly points of knowledge exchange between citing and cited publications. The proposed citation weight places a premium on allocating weight based on the uniqueness of contribution from every knowledge exchange point (citation mentions). In theory, the proposed weight is a count of unique contribution from the cited and citing article, given that zero weight is added whenever a citation mention is not different from the others; therefore, a more nuanced method than the number of citations and number of citation mentions.

It is interesting to note that the proposed semantic similarity-based citation weight correlated highly with the number of citations. A partial list of metrics that highly corelated with number of citations in previous studies include citation mentions, number of multiple citation mentions, the sum of multiple citation mentions (N-weighted recitation weight) and N-squared citation counting (Zhao & Strotmann, 2016). In contrast, Hassan et al. (2017) reported a weak correlation coefficient between the importance of a paper and the average number of citations received per year. Similarly, the number of citations, while regarded as a weak feature in this previous study, was yet classified as one of the best features for classifying citations based on importance (Valenzuela et al., 2015b)

Although the relationship between the number of citations and the proposed citation weight was high and statistically significant, the comparison of the ranks (see Appendix B) of the sampled articles by the proposed citation weight and citation number shows only 7% of the sampled articles retained their ranks. Of the 93% of the sampled articles that changed ranks, seven articles changed at least 20 places, and the most significant change in position is 30. The average position change was 8.28. This result shows that though the proposed citation weight is statistically related to citation number, the change in ranks of the sampled articles using the two metrics shows that they are practically different metrics.

Among other metrics, the relationship between the proposed citation weights and citation mentions is the strongest statistically. Notably, citation mentions had a perfect correlation with the proposed semantic similarity-based citation weights. It is speculated that the high number of single mentions which constituted about 73% of the sample contributed to the high correlation. First, it appears that the effect of discounting citation count in favour of placing more weights on uniqueness was unexpectedly not visible on the relationship between the two metrics. Secondly, the high correlation between the proposed semantic similarity-based citation weight and the number of citation mentions suggests that the proposed citation weights may rank among important metrics for determining the contribution of citations. Previous studies regarded the number of citation mentions as the strongest feature for measuring the contribution of cited documents in the citing documents (Yu et al., 2019; Zhu et al., 2015). In light of this, it will be of interest to find out in future studies how well the proposed citation weight performs in predicting the contribution, importance or influence of cited publications in citing publications.

It is important to note that the strong relationship between the number of citation mentions and the proposed citation weighting connotes the importance of the proposed citation weighting in predicting and measuring contribution. The proposed semantic similarity-based citation weighting, which is a derivative of citation mention, could be categorized as a contribution metric, given that Yu et al. (2019) suggested that the number of citation mentions is a useful metric for measuring contribution. Similarly, Zhu et al. (2015) found in-text citation mention as the most vital feature for predicting scientific publications' academic influence. Zhao et al. (2017) mapped uni-citation and multi-citations to citation functions. It was observed that most single mentioned citations were non-essential (either perfunctory or reviewed citation functions), while multiple citations were likely to be influential on the citing paper.

It appears despite the high likelihood of citation mentions to predict the importance or contribution of publications and its high correlation with content-based metrics that predict contribution, content-based metrics perform better in some studies in predicting the contribution of cited publications than the syntactic feature. In Hassan et al. (2017), citation mentions had the highest correlation coefficient among other contextual, cue word-based, and textual features for determining important citation classes. However, Hassan et al. (2017) further found out that the correlation coefficient did not tell the whole story as features such as the similarity between the abstract of cited paper and text of citing paper, and cue words for using and extending research were more informative in classifying citations into important/non-important classes than the number of citation mentions. Similarly, for Yu et al. (2019), result was mixed; despite the usefulness of citation mention frequency for measuring contribution, the correlation between relevancy, a content-based metric, and the number of citation mention was not linear. Yu et al. (2019) regarded "relevance" as a degree of contribution from the cited article to the citing article. The statistical relationship between relevance and citation mentions was low but was the strongest among other investigated metrics.

Unsurprising, the change in rank when the sampled articles were ranked using the proposed citation weights and citation mentions is the lowest among other investigated metrics, showing that these two metrics are the most similar. The average rank change was 2.08, while 77% changed ranks, and the highest rank change was 12.

Surprisingly, the relationship between the proposed citation weights and each of two other metrics, the number of multiple citation mentions and the sum of multiple citation mentions was moderately strong and significant. The two metrics had the same correlation coefficient with the proposed citation weights. This result implies that the removal of single mentioned citations (supposedly the non-essential citations) does not improve the relationship between citation mentions and the proposed metric for measuring contribution, as revealed earlier that citation mention count had a higher correlation coefficient than the multiple citation metrics. Secondly, given that the two metrics under consideration had the same value of correlation co-efficient with the proposed semantic similarity-based citation weights suggests that counting or summing multiple citation mentions makes no difference in practice for the prediction of contribution.

The sampled 100 articles, when ranked with the number of multiple citation mentions, are presented in

Appendix ***E***. The average rank change was 9.66, while 97% of the sampled articles changed ranks, and the highest rank change was 35. The result of the sampled 100 articles when ranked with the sum of multiple citation mentions is presented in

Appendix ***F***. The change in rank when the sampled articles were ranked by the proposed citation weights and sum of citation mentions was the highest of the four existing metrics after the number of positive citation sentiments. The average rank change was 10.52, while 94% of the sampled articles changed ranks, and the highest rank change was 35.

The correlation coefficient of the the proposed semantic similarity-based citation and the number of positive citation sentiments, though moderately strong, positive and significant, was the lowest among the metrics that were compared with the proposed metric. Yan et al., (2020) found a positive, weak and significant correlation between sentiment and journal citation impact. The result of the sampled 100 articles when ranked with the number of positive citations is presented in Appendix G. The change in rank when the sampled articles were ranked by the proposed citation weights and the number of positive citations was the highest. The average rank change was 11.82, while 96% of the sampled articles changed ranks, and the highest rank change was 56.

5.2 Research Question 2: How different is the proposed semantic similarity-based residual 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 generation 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 generation 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.

5.3 Research Question 3: What differences exist in the residual citations 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 number of residual citations due to an article reduces as the number of citation generations also increases. Like the cascading citation weighting system, the average amount of residual citations accrued from the indirect citation was highest at the second generation and reduced as the number of generations 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 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.

5.4 Research Question 4: What is the Residual Citation pattern between 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 onto the number 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 are 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. The first category includes 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 categories 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 numbers 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.

Chapter 6

6 Conclusion and Recommendation

6.1 Summary and Conclusion

6.1.1 Direct Citation Weighting

This thesis proposes a weighting method for citation mentions/contexts, where unique citation contexts are allocated more weights. This thesis is built on the citation mention analysis as a method for weighting citations. Like citation mention analysis, cited articles with more citation mentions receive more weights. A total of 100 publications that received moderate citations (maximum=179, minimum=31, average=73.17) and were published in 2014 were systematically sampled from the Web of Science database. Most of the articles (73.02%) were mentioned once in the citing articles. There were 11,234 citation contexts; 5918 citations were mentioned more than once in the citing publications. A total of 9795 citation context pairs were obtained from the 5918 citation contexts that were mentioned more once. Citation context pairing occurred between at least two citation mentions of a cited document in the citing document. All the possible citation context pairs in multiple citation mentions were obtained. The number of citation context pairs per citing document was obtained as n combination 2 ($^{n}C_{2}$), where n is the number of citation mentions.

A python program that implemented the BioSentVec sentence embeddings model was used for automating the semantic similarity measurement between citation context pairs. The semantic similarity score between a citation context pair represented as the cosine value of the angle between the representations of the citation contexts on vector space was a numerical value between zero and one. Manually classified citation context pairs to three semantic similarity classes (similar, somewhat similar, and not similar) was necessary since there were no existing human-annotated citation context semantic similarity corpora. Out of the 9795 pairs of citation contexts, a total of 981 citation context pairs were given to two biomedical experts for annotation. The annotation was used for creating boundaries between zero and one for the three semantic similarity classes, semantic similarity score equal to zero and less than 0.51 were categorized as “not similar” and allocated weight of one, semantic similarity scores that were equal to 0.51 and less than 0.71, were categorized as “somewhat similar” and allocated weight of 0.5, while semantic similarity scores that fell between to 0.71 and one were categorized as “similar” and allocated weights of zero.

Spearman rho’s correlation test revealed citation mentions was the most similar metric to the proposed direct citation metric. This result implies the expected impact of discounting the ordinary count of citation mentions was of no effect on the ranking of the sampled publication using the proposed citation weighting system and the number of citation mentions. Other existing metrics were less similar to the proposed citation weighting system, and this was reflected in the rank change, which varied significantly when the sampled articles were ranked using the proposed citation weighting in comparison to the existing metrics.

6.1.2 Indirect Citation Weighting

Indirect citation weighting is concerned with allocating weights that are based on the contribution of a previously cited article in publications that did not cite it, but exist as a generation of citation on its citation path. Allocating weights to indirect citation in this thesis is exploratory and the research inquiries focused on the following; are there situations where papers should be allocated residual citations from papers that indirectly cited them by considering the contribution of the previously cited paper in the generations of its citation.

The top ten most cited biomedical publications that were published in 2014 were sampled as the base articles. A total of 50 first-generation articles, 100 second-generation articles, 200 third-generation articles, 400 fourth generation articles and 710 fifth-generation articles were sampled. Citation context pairs were obtained from the citation context of the first generation articles and citation contexts of nth generation citation. Citation context pairs were fed to the already trained python program that is based on the already trained BioSentVec model to obtain the semantic similarity scores. The citation context pair of a first-generation article and its nth generation article with the highest semantic similarity score was selected as proof of the contribution of base article in the the contribution of the base article in the nth generation article and considered for allocating residual citation. The selected semantic similarity scores were classified as “not similar”, “somewhat similar”, and “similar” using the boundaries obtained in the direct citation weighting aspect of this thesis.

The result of the indirect citation weighting part of this study revealed like the cascading citation system, that residual citations to articles from their generations of citations decreased as the number of generations increased. However, residual citations accrued to publications at all the generations were statistically different between the proposed indirect weighting and the cascading citation system.

6.2 Recommendations

According to the results of the analysis of the collected data, the proposed semantic similarity-based citation weighting method is similar to citation mention frequency weight. However, it is recommended that the proposed method be used in future studies with bigger datasets and improved computational methods for obtaining semantic similarity in biomedical and other fields. It is obvious that obtaining the number of citation mentions is simpler than the proposed method; however, the proposed method presents an alternative method for weighting the contribution of cited documents in the citing documents.

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 consequent 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.

6.3 Contributions of the Thesis

This study contributed both to practice and research in Computer Science and Library Information Science disciplines, specifically scholarly communication, bibliometrics, data science, and natural language processing sub-fields.

6.3.1 Datasets

One of the research outputs of this thesis is the datasets from the direct and indirect citation weighting parts. There were 11,234 citation contexts for the direct citation and 1,257 citation contexts for the indirect citation aspects of this doctoral thesis, which translated to 9,795 citation context pairs and 5,272 citation context pairs, respectively. Citation context datasets were collected manually from the sampled articles and without a pre-determined window of words or sentences. Similarly to the result in earlier studies, 73.02% of the sampled articles' citations for the direct citation context weighting were mentioned once in the citing articles. An analysis of the full texts of the journal of informetrics', a LIS journal, 74.3% were cited once (Hu et al., 2015). In a multidisciplinary study by Boyack et al. (2018), 71.5% of the in-text citations from PubMed Central Open Access Subset (PCMOA) and 69.5% of the in-text citations from Elsevier (ELS) journals were cited once. Given that the proposed semantic similarity-based citation weighting method is useful for weighting multi-mentioned citation contexts, the practical implication of this observation is only about 25% of citations can be qualified for weighting.

One of the merits of the datasets is the manual data collection method, without a fixed window of sentences or words. Data collection was focused on painstakingly identifying the span of text that represented the cited knowledge in a well-described manner (see section 3.2 for details). The data collection is unlike the method that is used to identify citation context in recent large scale studies, which adopts a pre-determined window of words or sentences and is optimized for collecting large corpora. A premium was placed on accurately identifying the citation context as opposed to quantity; therefore, strict steps were followed to achieve this goal. Identifying citation contexts accurately without a pre-determined window of text is a work-in-progress(Kang & Kim, 2012; Kaplan et al., 2009; Ou & Kim, 2018). Recent studies relied on the use of a fixed window of texts for automatic (Houngbo & Mercer, 2017; Singha Roy et al., 2020) and manual (Tabatabaei, 2013) identification of citation contexts and this method has been reported in recent studies. While automating the use of a fixed window of texts have led to the successful collection of large corpora in many studies, research has shown that pre-determined window of texts does not always accurately represent the citation contexts (Cohan et al., 2015; Kaplan et al., 2009; Ritchie et al., 2006). In contrast to the existing method of collecting citation context data using fixed windows of words or sentences, fairly large citation context corpora were collected in this study.

Another dataset from this thesis is the annotated sample of about 10.02% of the direct citation dataset. This dataset had 981 citation context pairs that were annotated into three semantic similarity classes -similar, somewhat similar and not similar- by two human experts. The sampled datasets were annotated to identify the thresholds between the three semantic similarity classes, and there was no corpus in the literature with citation contexts in the three semantic similarity classes. The two experts independently annotated all the sampled 981 citation context pairs, and the inter-coder agreement of the two coders was 66.16% based on the percentage agreement and 0.27 Cohen Kappa score. The inter-rater reliability was fair; better inter-rater reliability is desired.

6.3.2 New Bibliometric Measures

One of the contributions of this thesis are the proposed metrics-semantic similarity-based citation weighting method and the residual citation weighting method. The proposed metrics are based on methods that are different from previous metrics, though the results suggest that the proposed direct citation metric is similar to citation mention count, unlike the residual citation metric that is different from the existing cascading citation method.

6.3.3 Relatedness on Citation Path

The proposed metrics will potentially start new discussion in scholarly communications discipline. For instance, new discussion about the idea of residual citation may lead to a new area of research interest because of the potentials for acknowledging contributions beyond the conventional direct citation as it is currently known. Secondly, the residual citation method exposes new way of exploring relatedness on citation networks, which can impact bibliometric enhance-information retrieval.

6.4 Limitations of the Study

The experts that were employed to annotate a sample of the citation contexts were early career professionals as their knowledge of the discipline is budding. Employing more experienced professionals for the data annotation would have been more desirable. Furthermore, the inter-rater agreement was below 70%, while the Cohen Kappa score was not strong (0.27). This is a limitation because lower inter-rater reliability scores connotes lower reliability of data collection instrument. However, the effect of the lower-inter coder reliability was tempered by considering only the datapoints where the two annotators agreed, while other records where they disagreed were disregarded. Besides, coding such a diverse and multidisciplinary texts may be more difficult than texts from narrower or more specific disciplines/sub-discipline such as Biochemistry, Chemistry, Library and Information Science etc. This could be one of explanations for the low inter-rater agreement between the expert annotators.

Secondly, only biomedical publications were considered for this study. This may affect the external validity of the result of this thesis. Perhaps the results would be different if corpus from other disciplines were considered where citing practices are different. Similarly, half-life of publications can play a role in the residual citation pattern of publications. In fields where there are short half-lives, subsequent generations of residual citation characteristics could vary a lot.

6.5 Suggestions for Further Studies

It is suggested that future studies consider to investigate:

1. the relationship between citation context polarity, number of citation mentions, and function on the contribution of a publication in the generations of its citation,
2. the correlation between the proposed semantic similarity-based citation weighting and citation polarity, citation importance and citation function.

3. the performance of the proposed weighting methods using bigger datasets in different disciplines, and
4. the performance of the proposed weighting methods using different methods for obtaining semantic similarity scores with annotated datasets by more experienced professionals and more trainings.

References

Abu-Jbara, A., Ezra, J., & Radev, D. (2013). Purpose and Polarity of Citation: Towards NLP-based Bibliometrics. *Proceedings of NAACL-HLT 2013*, 596–606.

Asubiaro, T. V. (2018). Research Collaboration Landscape of the University of Ibadan Biomedical Authors between 2006 and 2015. *African Journal of Library, Archives and Information Science*, *28*(1), 15.

Athar, A. (2011). Sentiment Analysis of Citations using Sentence Structure-Based Features. *Proceedings of the ACL-HLT 2011 Student Session*, 81–87.

Athar, A., & Teufel, S. (2012a). Context-Enhanced Citation Sentiment Detection. *2012 Conference of the North American Chapter of the Association for Computational Linguistics*, 597–601.

Athar, A., & Teufel, S. (2012b). Detection of Implicit Citations for Sentiment Detection. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 18–26.

Baumgartner, H., & Pieters, R. (2003). The structural influence of marketing journals: A citation analysis of the discipline and its subareas over time. *Journal of Marketing*, *67*(2), 123–139.

Bennett, D. M., & Taylor, D. M. (2003). Unethical practices in authorship of scientific papers. *Emergency Medicine Australasia*, *15*(3), 263–270.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The EigenfactorTM Metrics. *Journal of Neuroscience*, *28*(45), 11433–11434. https://doi.org/10.1523/JNEUROSCI.0003-08.2008

Biscaro, C., & Giupponi, C. (2014). Co-Authorship and Bibliographic Coupling Network Effects on Citations. *PLoS ONE*, *9*(6), e99502. https://doi.org/10.1371/journal.pone.0099502

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *ArXiv:1607.04606 [Cs]*. http://arxiv.org/abs/1607.04606

Bonzi, S. (1982). Characteristics of a Literature as Predictors of Relatedness Between Cited and Citing Works. *Journal of the American Society for Information Science*, *33*(4), 208–216.

Bornmann, L., & Leydesdorff, L. (2014). Scientometrics in a changing research landscape: Bibliometrics has become an integral part of research quality evaluation and has been changing the practice of research. *EMBO Reports*, *15*(12), 1228–1232. https://doi.org/10.15252/embr.201439608

Boyack, K. W., van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, *12*(1), 59–73. https://doi.org/10.1016/j.joi.2017.11.005

Cano, V. (1989). Citation behavior: Classification, utility, and location. *Journal of the American Society for Information Science and Technology*, *40*(4), 284–290.

Caragea, C., Bulgarov, F. A., Godea, A., & Das Gollapalli, S. (2014). Citation-Enhanced Keyphrase Extraction from Research Papers: A Supervised Approach. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1435–1446. https://doi.org/10.3115/v1/D14-1150

Chen, Q., Peng, Y., & Lu, Z. (2019). BioSentVec: Creating sentence embeddings for biomedical texts. *The Seventh IEEE International Conference on Healthcare Informatics*, 5. https://arxiv.org/abs/1810.09302

Cohan, A., Ammar, W., van Zuylen, M., & Cady, F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. *Proceedings of the 2019 Conference of the North*, 3586–3596. https://doi.org/10.18653/v1/N19-1361

Cohan, A., Soldaini, L., & Goharian, N. (2015). Matching Citation Text and Cited Spans in Biomedical Literature: A Search-Oriented Approach. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1042–1048. https://doi.org/10.3115/v1/N15-1110

Colledge, L. (2014). *Snowball Metrics Recipe Book.* Elsevier. https://www.snowballmetrics.com/wp-content/uploads/snowball-recipe-book_HR.pdf

Corbyn, Z. (2008). *Researchers may play dirty to beat REF*. Times Higher Education. https://www.timeshighereducation.com/news/researchers-may-play-dirty-to-beat-ref/400516.article

Cottrill, C. A., Rogers, E. M., & Mills, T. (1989). Co-citation Analysis of the Scientific Literature of Innovation Research Traditions: Diffusion of Innovations and Technology Transfer. *Knowledge*, *11*(2), 181–208. https://doi.org/10.1177/107554708901100204

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

Di Marco, C., Kroon, F., & Mercer, R. (2006). Using Hedges to Classify Citations in Scientific Articles. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (Vol. 20, pp. 247–263). Springer. https://doi.org/10.1007/1-4020-4102-0_19

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.

Dong, C., & Schafer, U. (2011). Ensemble-style Self-training on Citation Classification. *Proceedings of 5th International Joint Conference on Natural Language Processing*, 623–631.

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

Fiala, D. (2012). Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, *6*(3), 370–388. https://doi.org/10.1016/j.joi.2012.02.002

Fiala, D., Šubelj, L., Žitnik, S., & Bajec, M. (2015). Do PageRank-based author rankings outperform simple citation counts? *Journal of Informetrics*, *9*(2), 334–348. https://doi.org/10.1016/j.joi.2015.02.008

Fiala, D., & Tutoky, G. (2017). PageRank-based prediction of award-winning researchers and the impact of citations. *Journal of Informetrics*, *11*(4), 1044–1068. https://doi.org/10.1016/j.joi.2017.09.008

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLOS ONE*, *12*(12), e0187394. https://doi.org/10.1371/journal.pone.0187394

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

Fujimagari, H., & Fujita, K. (2014). *Detecting Research Fronts Using Neural Network Model for Weighted Citation Network Analysis*. 131–136. https://doi.org/10.1109/IIAI-AAI.2014.36

Garzone, M., & Mercer, R. E. (2000). Towards an Automated Citation Classifier. In H. J. Hamilton (Ed.), *Advances in Artificial Intelligence* (Vol. 1822, pp. 337–346). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45486-1_28

Glänzel, W., & Schubert, A. (2004). Chapter 11: Analysing Scientific Networks through co-authorship. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 257–276). Kluwer Acad. Publ.

Hassan, S.-U., Akram, A., & Haddawy, P. (2017). Identifying Important Citations Using Contextual Information from Full Text. *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference On*, 1–8.

Hassan, S.-U., Safder, I., Akram, A., & Kamiran, F. (2018). A novel machine-learning approach to measuring scientific knowledge flows using citation context analysis. *Scientometrics*, *116*(2), 973–996. https://doi.org/10.1007/s11192-018-2767-x

Herlach, G. (1976). Can Retrieval of Information from Citation Indexes be Simplified? Multiple Mention of a Reference as a Characteristic of the Link between Cited and Citing Article. *Journal of the American Society for Information Science*, *29*(6), 308.

HernáNdez-Alvarez, M., & Gomez, J. M. (2016). Survey about citation context analysis: Tasks, techniques, and resources. *Natural Language Engineering*, *22*(03), 327–349. https://doi.org/10.1017/S1351324915000388

Houngbo, H., & Mercer, R. E. (2017). Investigating Citation Linkage with Machine Learning. *Advances in Artificial Intelligence*, 78–83. https://doi.org/10.1007/978-3-319-57351-9_10

Hu, Z., Chen, C., & Liu, Z. (2015). The Recurrence of Citations within a Scientific Article. *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference*, 221–229.

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

Jeong, Y. (2016, March 10). *Applying Content-based Similarity Measure to Author Co-citation Analysis*. https://doi.org/10.9776/16212

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

Jurgens, D., Kumar, S., Hoover, R., McFarland, D., & Jurafsky, D. (2016). Citation Classification for Behavioral Analysis of a Scientific Field. *ArXiv:1609.00435 [Cs]*. http://arxiv.org/abs/1609.00435

Kang, I.-S., & Kim, B.-K. (2012). Characteristics of Citation Scopes: A Preliminary Study to Detect Citing Sentences. In T. Kim, J. Ma, W. Fang, Y. Zhang, & A. Cuzzocrea (Eds.), *Computer Applications for Database, Education, and Ubiquitous Computing* (pp. 80–85). Springer. https://doi.org/10.1007/978-3-642-35603-2_11

Kaplan, D., Iida, R., & Tokunaga, T. (2009). Automatic Extraction of Citation Contexts for Research Paper Summarization: A Coreference-chain based Approach. *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009*, 88–95.

Katz, S., & Martin, B. (1997). What is Research Collaboration. *Research Policy*, *26*, 1–18.

Khalid, A., Alam, F., & Ahmed, I. (2018). Extracting reference text from citation contexts. *Cluster Computing*, *21*(1), 605–622. https://doi.org/10.1007/s10586-017-0954-9

Kim, H. J., Jeong, Y. K., & Song, M. (2016). Content- and proximity-based author co-citation analysis using citation sentences. *Journal of Informetrics*, *10*(4), 954–966. https://doi.org/10.1016/j.joi.2016.07.007

Knoth, P., & Herrmannova, D. (2014). Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication's Contribution. *D-Lib Magazine*, *20*(11/12). https://doi.org/10.1045/november2014-knoth

Larivière, V., Sugimoto, C. R., & Cronin, B. (2012). A bibliometric chronicling of library and information science's first hundred years. *Journal of the American Society for*

Information Science and Technology, *63*(5), 997–1016. https://doi.org/10.1002/asi.22645

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology*, *60*(7), 1327–1336. https://doi.org/10.1002/asi.21024

Li, X., He, Y., Meyers, A., & Grishman, R. (2013). Towards Fine-grained Citation Function Classification. *Proceedings of Recent Advances in Natural Language Processing*, 402–407.

Liu, S., Chen, C., Ding, K., Wang, B., Xu, K., & Lin, Y. (2014). Literature retrieval based on citation context. *Scientometrics*, *101*(2), 1293–1307. https://doi.org/10.1007/s11192-014-1233-7

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

Manning, C., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

Maričić, S., Spaventi, J., Pavičić, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the American Society for*

Information Science, *49*(6), 530–540. https://doi.org/10.1002/(SICI)1097-4571(19980501)49:6<530::AID-ASI5>3.0.CO;2-8

Maričić, S., Spaventi, J., Pavičić, L., & Pifat-Mrzljak, G. (1998). Citation context versus the frequency counts of citation histories. *Journal of the American Society for Information Science*, *49*(6), 530–540. https://doi.org/10.1002/(SICI)1097-4571(19980501)49:6<530::AID-ASI5>3.0.CO;2-U

McCain, K. W., & Turner, K. (1989). Citation context analysis and aging patterns of journal articles in molecular genetics. *Scientometrics*, *17*(1–2), 127–163. https://doi.org/10.1007/BF02017729

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

Meng, R., Lu, W., Chui, Y., & Shuguang, H. (2017). Automatic Classification of Citation Function by New Linguistic Features. *IConference 2017 Proceedings*, 826–830. https://doi.org/10.9776/17349

Mercer, R. E., Di Marco, C., & Kroon, F. W. (2004). The Frequency of Hedging Cues in Citation Contexts in Scientific Writing. In A. Y. Tawfik & S. D. Goodwin (Eds.), *Advances in Artificial Intelligence* (Vol. 3060, pp. 75–88). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-24840-8_6

Meyers, A. (2013). Contrasting and Corroborating Citations in Journal Articles. *Proceedings of Recent Advances in Natural Language Processing*, 460–466.

Moravcsik, M., & Murugesan, P. (1975). Some Results on the Function and Quality of Citations. *Social Studies of Science*, *5*(1), 86–92.

Nazir, S., Asif, M., Ahmad, S., Bukhari, F., Afzal, M. T., & Aljuaid, H. (2020). Important citation identification by exploiting content and section-wise in-text citation count. *PLOS ONE*, *15*(3), e0228885. https://doi.org/10.1371/journal.pone.0228885

Ou, S., & Kim, H. (2018). Unsupervised Citation Sentence Identification Based on Similarity Measurement. In G. Chowdhury, J. McLeod, V. Gillet, & P. Willett (Eds.), *Transforming Digital Worlds* (pp. 384–394). Springer International Publishing. https://doi.org/10.1007/978-3-319-78105-1_42

Page, L., Sergey, B., Rajeev, M., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web* (SIDL-WP-1999-0120). Stanford University, CA, USA.

Peritz, B. C. (1983). A classification of citation roles for the social sciences and related fields. *Scientometrics*, *5*(5), 303–312. https://doi.org/10.1007/BF02147226

Prathap, G., & Nishy, P. (2016). An Alternative Size-Independent Journal Performance Indicator for Science on the Periphery. *Current Science*, *111*(11), 1802. https://doi.org/10.18520/cs/v111/i11/1802-1810

Prathap, G., Nishy, P., & Savithri, S. (2016). On the Orthogonality of Indicators of Journal Performance. *Current Science*, *111*(5), 876. https://doi.org/10.18520/cs/v111/i5/876-881

Pride, D., & Knoth, P. (2017). *Incidental or influential?–A decade of using text-mining for citation function classification.* 16th International Society of Scientometrics and Informetrics Conference, Wuhan, China.

Qayyum, F., & Afzal, M. T. (2019). Identification of important citations by exploiting research articles' metadata and cue-terms from content. *Scientometrics*, *118*(1), 21–43. https://doi.org/10.1007/s11192-018-2961-x

Ritchie, A., Robertson, S., & Teufel, S. (2008). Comparing citation contexts for information retrieval. *Proceeding of the 17th ACM Conference on Information and Knowledge Mining - CIKM '08*, 213. https://doi.org/10.1145/1458082.1458113

Ritchie, A., Teufel, S., & Robertson, S. (2006). How to Find Better Index Terms Through Citations. *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval?*, 25–32. https://doi.org/10.3115/1629808.1629813

Sellers, S. L., Mathiesen, S. G., Perry, R., & Smith, T. (2004). Evaluation of Social Work Journal Quality: Citation versus Reputation Approaches. *Journal of Social Work Education*, *40*(1), 143–160. https://doi.org/10.1080/10437797.2004.10778484

Singha Roy, S., Mercer, R. E., & Urra, F. (2020). Investigating Citation Linkage as a Sentence Similarity Measurement Task Using Deep Learning. In C. Goutte & X.

Zhu (Eds.), *Advances in Artificial Intelligence* (pp. 483–495). Springer International Publishing. https://doi.org/10.1007/978-3-030-47358-7_50

Smedt, T., & Daelemans, W. (2012). Pattern for Python. *Journal of Machine Learning Research*, *13*, 2063–2067.

Soğancıoğlu, G., Öztürk, H., & Özgür, A. (2017). BIOSSES: A semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, *33*(14), i49–i58. https://doi.org/10.1093/bioinformatics/btx238

Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. *International Journal of Research in Marketing*, *32*(1), 64–77. https://doi.org/10.1016/j.ijresmar.2014.09.004

Strotmann, A., & Zhao, D. (2014). Uncertainty of author citation rankings: Lessons from in-text citation weighing schemes. *Proceedings of the Association for Information Science and Technology*, *51*(1), 1–4.

Sun, X., Ding, K., & Lin, Y. (2016). Mapping the evolution of scientific fields based on cross-field authors. *Journal of Informetrics*, *10*(3), 750–761. https://doi.org/10.1016/j.joi.2016.04.016

Tabatabaei, N. (2013). *Contribution of Information Science to Other Disciplines as Reflected in Citation Contexts of Highly Cited JASIST Papers* [PhD Thesis]. McGill University.

Teixeira da Silva, J. A., & Dobránszki, J. (2016). Multiple Authorship in Scientific Manuscripts: Ethical Challenges, Ghost and Guest/Gift Authorship, and the Cultural/Disciplinary Perspective. *Science and Engineering Ethics*, *22*(5), 1457–1472. https://doi.org/10.1007/s11948-015-9716-3

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing - EMNLP '06*, 103. https://doi.org/10.3115/1610075.1610091

Tuarob, S., Kang, S. W., Wettayakorn, P., Pornprasit, C., Sachati, T., Hassan, S.-U., & Haddawy, P. (2020). Automatic Classification of Algorithm Citation Functions in Scientific Literature. *IEEE Transactions on Knowledge and Data Engineering*, *32*(10), 1881–1896. https://doi.org/10.1109/TKDE.2019.2913376

Turney, P. D., & Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. https://doi.org/10.1613/jair.2934

Valenzuela, M., Ha, V., & Etzioni, O. (2015a). Identifying Meaningful Citations. *AAAI Workshops*, 21–26.

Valenzuela, M., Ha, V., & Etzioni, O. (2015b). Identifying Meaningful Citations. *Scholarly Big Data: AI Perspectives, Challenges, and Ideas*, 6.

Voos, H., & Dagaev, K. (1976). Are All Citations Equal? Or, Did We op. Cit. Your Idem? *The Journal of Academic Librarianship*, *1*(6), 19–21.

Wagner, C. S., & Leydesdorff, L. (2005). Mapping the network of global science: Comparing international co-authorships from 1990 to 2000. *International Journal of Technology and Globalisation*, *1*(2), 185–208.

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

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

Wang, M., Zhang, J., Jiao, S., Zhang, X., Zhu, N., & Chen, G. (2020). Important citation identification by exploiting the syntactic and contextual information of citations. *Scientometrics*, *125*(3), 2109–2129. https://doi.org/10.1007/s11192-020-03677-1

Wang, Y., Afzal, N., Fu, S., Wang, L., Shen, F., Rastegar-Mojarad, M., & Liu, H. (2018). MedSTS: A resource for clinical semantic textual similarity. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-018-9431-1

Wilhite, A. W., & Fong, E. A. (2012). Coercive Citation in Academic Publishing. *Science*, *335*(6068), 542–543. https://doi.org/10.1126/science.1212540

Xu, J., Zhang, Y., Wu, Y., Wang, J., Dong, X., & Xu, H. (2015). Citation sentiment analysis in clinical trial papers. *AMIA Annual Symposium Proceedings*, *2015*, 1334.

Yan, E., Chen, Z., & Li, K. (2020). The relationship between journal citation impact and citation sentiment: A study of 32 million citances in PubMed Central. *Quantitative Science Studies*, *1*(2), 664–674. https://doi.org/10.1162/qss_a_00040

Yan, E., & Ding, Y. (2010). Weighted citation: An indicator of an article's prestige. *Journal of the American Society for Information Science and Technology*, *61*(8), 1635–1643. https://doi.org/10.1002/asi.21349

Yan, E., Ding, Y., & Sugimoto, C. R. (2011). P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, *62*(3), 467–477. https://doi.org/10.1002/asi.21461

Yu, G., Wang, W., Pak, C., & Yu, T. (2019). Research on the Relevancy of Scientific Literature Based on the Citation-Mention Frequency. *IEEE Access*, *7*, 181750–181757. https://doi.org/10.1109/ACCESS.2019.2958952

Yu, T., Yu, G., & Wang, M.-Y. (2014). Classification method for detecting coercive self-citation in journals. *Journal of Informetrics*, *8*(1), 123–135. https://doi.org/10.1016/j.joi.2013.11.001

Zhang, C.-T. (2009). A proposal for calculating weighted citations based on author rank. *EMBO Reports*, *10*(5), 416–417. https://doi.org/10.1038/embor.2009.74

Zhao, D., Cappello, A., & Johnston, L. (2017). Functions of Uni- and Multi-citations: Implications for Weighted Citation Analysis. *Journal of Data and Information Science*, *2*(1), 51–69. https://doi.org/10.1515/jdis-2017-0003

Zhao, D., & Logan, E. (2002). Citation analysis using scientific publications on the Web as data source: A case study in the XML research area. *Scientometrics*, *54*(3), 449–472.

Zhao, D., & Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996-2005: Introducing author bibliographic-coupling analysis. *Journal of the American Society for Information Science and Technology*, *59*(13), 2070–2086. https://doi.org/10.1002/asi.20910

Zhao, D., & Strotmann, A. (2014a). Weighted in-text citations and research impact patterns: A case study of library and information science. *Proceedings of the American Society for Information Science and Technology*, *51*(1), 1–5. https://doi.org/10.1002/meet.2014.14505101097

Zhao, D., & Strotmann, A. (2014b). In-text author citation analysis: Feasibility, benefits, and limitations. *Journal of the Association for Information Science and Technology*, *65*(11), 2348–2358. https://doi.org/10.1002/asi.23107

Zhao, D., & Strotmann, A. (2016). Dimensions and Uncertainties of Author Citation Rankings: Lessons Learned From Frequency-Weighted In-Text Citation Counting. *Journal of the Association for Information Science and Technology*, *67*(3), 671–682. https://doi.org/10.1002/asi.23418

Zhao, D., & Strotmann, A. (2015). Re-citation Analysis: A Promising Method for Improving Citation Analysis for Research Evaluation, Knowledge Network Analysis, Knowledge Representation and Information Retrieval. *Proceedings of the 15th International Conference of the International Society for Scientometrics and Informetrics*, 1061–1065.

Zhao, D., Strotmann, A., & Cappello, A. (2018). In-text function of author self-citations: Implications for research evaluation practice. *Journal of the Association for Information Science and Technology*, *69*(7), 949–952. https://doi.org/10.1002/asi.24046

Zhu, X., Turney, P., Lemire, D., & Vellino, A. (2015). Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, *66*(2), 408–427. https://doi.org/10.1002/asi.23179

Appendices

Appendix 1

Appendix A: Sampled 100 articles for the Direct Citation aspect of this Thesis

WoS rank	Title	actual	# of pairs	# of citations from WoS	removed
Article 1	Right heart dysfunction in heart failure with preserved ejection fraction.	180	349	200	21
Article 2	The application of esophageal pressure measurement in patients with respiratory failure.	162	84	180	13
Article 3	Combined photothermal and photodynamic therapy delivered by PEGylated MoS2 nanosheets.	167	218	188	26
Article 4	Development of the EUCAST disk diffusion antimicrobial susceptibility testing method and its implementation in routine microbiology laboratories.	157	48	170	13
Article 5	Structural basis for Ca2+ selectivity of a voltage-gated calcium channel.	137	90	155	6
Article 6	Highly conducting, strong nanocomposites based on nanocellulose-assisted aqueous dispersions of single-wall carbon nanotubes.	149	298	157	20
Article 7	Consensus statement on the diagnosis, treatment and follow-up of patients with primary adrenal insufficiency.	110	425	140	12
Article 8	ER contact sites define the position and timing of endosome fission.	124	170	142	18
Article 9	Complementary genomic approaches highlight the PI3K/mTOR pathway as a common vulnerability in osteosarcoma.	128	67	136	18

Article 10	Temporal trends in marijuana attitudes, availability and use in Colorado compared to non-medical marijuana states: 2003-11.	118	119	130	13
Article 11	Front-line transplantation program with lenalidomide, bortezomib, and dexamethasone combination as induction and consolidation followed by lenalidomide maintenance in patients with multiple myeloma: a phase II study by the Intergroupe Francophone du Myelome.	101	642	150	40
Article 12	New materials graphyne, graphdiyne, graphone, and graphane: review of properties, synthesis, and application in nanotechnology.	117	107	115	5
Article 13	Pore size effect of collagen scaffolds on cartilage regeneration.	107	59	123	16
Article 14	Altitudinal changes in malaria incidence in highlands of Ethiopia and Colombia.	100	52	109	2
Article 15	Effects of dexamethasone as a local anaesthetic adjuvant for brachial plexus block: a systematic review and meta-analysis of randomized trials.	102	153	116	14
Article 16	Effects of resveratrol on gut microbiota and fat storage in a mouse model with high-fat-induced obesity.	110	211	107	5
Article 17	Use of epigenetic drugs in disease: an overview.	93	166	132	31
Article 18	Synucleins regulate the kinetics of synaptic vesicle endocytosis.	107	50	115	15
Article 19	Multiple APOBEC3 restriction factors for HIV-1 and one Vif to rule them all.	100	86	109	9
Article 20	Interactions of aluminum with biochars and oxidized biochars: implications for the biochar aging process.	102	138	102	5
Article 21	Ultrafast thin-disk laser with 80 muJ pulse energy and 242 W of average power.	68	279	101	5
Article 22	Platelets mediate lymphovenous hemostasis to maintain blood-lymphatic separation throughout life.	97	168	98	2
Article 23	The unfolded-protein-response sensor IRE-1alpha regulates the function of CD8alpha+ dendritic cells.	96	28	100	6
Article 24	Identification of pathways for bipolar disorder: a meta-analysis.	94	122	105	12

Article 25	Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein.	96	444	95	2
Article 26	Restoring visual function to blind mice with a photoswitch that exploits electrophysiological remodeling of retinal ganglion cells.	85	46	94	6
Article 27	A meta-analysis and review of holistic face processing.	93	152	91	4
Article 28	The liquid phase epitaxy approach for the successful construction of ultra-thin and defect-free ZIF-8 membranes: pure and mixed gas transport study.	88	340	95	10
Article 29	Oxygen at nanomolar levels reversibly suppresses process rates and gene expression in anammox and denitrification in the oxygen minimum zone off northern Chile.	87	21	88	3
Article 30	Synthesis of isatins by I2/TBHP mediated oxidation of indoles.	78	158	85	3
Article 31	Real-time contact force sensing for pulmonary vein isolation in the setting of paroxysmal atrial fibrillation: procedural and 1-year results.	69	43	83	3
Article 32	Organophosphorus flame retardants (PFRs) in human breast milk from several Asian countries.	85	39	82	2
Article 33	Community-supported models of care for people on HIV treatment in sub-Saharan Africa.	82	28	91	13
Article 34	Poly(3-hydroxybutyrate)/ZnO bionanocomposites with improved mechanical, barrier and antibacterial properties.	75	52	83	6
Article 35	Increasing sensing resolution with error correction.	77	131	81	6
Article 36	Recurrent glioblastoma treated with bevacizumab: contrast-enhanced T1-weighted subtraction maps improve tumor delineation and aid prediction of survival in a multicenter clinical trial.	80	31	80	6
Article 37	Multiplexed homogeneous assays of proteolytic activity using a smartphone and quantum dots.	80	15	78	4
Article 38	Defining language networks from resting-state fMRI for surgical planning--a feasibility study.	72	25	80	7

Article 39	Encapsulating Pd nanoparticles in double-shelled graphene@carbon hollow spheres for excellent chemical catalytic property.	73	14	77	4
Article 40	Enhanced colorimetric immunoassay accompanying with enzyme cascade amplification strategy for ultrasensitive detection of low-abundance protein.	74	136	79	7
Article 41	Body mass index categories and mortality risk in US adults: the effect of overweight and obesity on advancing death.	74	216	76	4
Article 42	High-flow nasal cannula versus conventional oxygen therapy after endotracheal extubation: a randomized crossover physiologic study.	58	37	72	1
Article 43	MiR-203 is downregulated in laryngeal squamous cell carcinoma and can suppress proliferation and induce apoptosis of tumours.	73	13	74	4
Article 44	Understanding trust as an essential element of trainee supervision and learning in the workplace.	72	78	89	20
Article 45	Molecular biomarkers in idiopathic pulmonary fibrosis.	52	74	105	37
Article 46	Pulsed-EPR evidence of a manganese(II) hydroxycarbonyl intermediate in the electrocatalytic reduction of carbon dioxide by a manganese bipyridyl derivative.	65	31	72	6
Article 47	Helicobacter pylori secreted peptidyl prolyl cis, trans-isomerase drives Th17 inflammation in gastric adenocarcinoma.	70	73	73	8
Article 48	Maintenance of postmitotic neuronal cell identity.	71	80	65	1
Article 49	Persistence of DNMT3A mutations at long-term remission in adult patients with AML.	66	12	67	3
Article 50	Microwave assisted extraction of pectin from waste Citrullus lanatus fruit rinds.	52	189	70	7
Article 51	Homochiral columns constructed by chiral self-sorting during supramolecular helical organization of hat-shaped molecules.	62	57	64	1
Article 52	Sleep, fatigue, depression, and circadian activity rhythms in women with breast cancer before and after treatment: a 1-year longitudinal study.	63	248	71	9
Article 53	Differential methylation of the TRPA1 promoter in pain sensitivity.	64	33	65	4

Article 54	MiR-429 inhibits cells growth and invasion and regulates EMT-related marker genes by targeting Onecut2 in colorectal carcinoma.	60	56	62	1
Article 55	Non-steroidal anti-inflammatory drug use in chronic pain conditions with special emphasis on the elderly and patients with relevant comorbidities: management and mitigation of risks and adverse effects.	50	27	68	8
Article 56	Evaluation of two matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) systems for the identification of Candida species.	52	42	66	7
Article 57	Additive-free hollow-structured Co3O4 nanoparticle Li-ion battery: the origins of irreversible capacity loss.	64	194	62	3
Article 58	Biology of adeno-associated viral vectors in the central nervous system.	61	433	78	20
Article 59	Locoregional recurrence risk for patients with T1,2 breast cancer with 1-3 positive lymph nodes treated with mastectomy and systemic treatment.	59	340	61	4
Article 60	Safety and efficacy of edoxaban, an oral factor Xa inhibitor, versus enoxaparin for thromboprophylaxis after total knee arthroplasty: the STARS E-3 trial.	43	106	59	2
Article 61	Breed differences in insulin sensitivity and insulinemic responses to oral glucose in horses and ponies of moderate body condition score.	39	68	58	2
Article 62	Biased agonism at G protein-coupled receptors: the promise and the challenges--a medicinal chemistry perspective.	63	12	59	3
Article 63	Pharmacogenetic-guided dosing of coumarin anticoagulants: algorithms for warfarin, acenocoumarol and phenprocoumon.	51	31	63	9
Article 64	Cavity quantum electrodynamics on a nanofiber using a composite photonic crystal cavity.	54	176	61	7
Article 65	Maternal diabetes and the risk of autism spectrum disorders in the offspring: a systematic review and meta-analysis.	50	30	61	7

Article 66	Interleukin-22 regulates the complement system to promote resistance against pathobionts after pathogen-induced intestinal damage.	61	59	59	5
Article 67	Attenuation and restoration of severe acute respiratory syndrome coronavirus mutant lacking 2'-o-methyltransferase activity.	59	19	58	4
Article 68	Coronavirus cell entry occurs through the endo-/lysosomal pathway in a proteolysis-dependent manner.	54	10	58	4
Article 69	Normal ranges of right ventricular systolic and diastolic strain measures in children: a systematic review and meta-analysis.	57	133	57	4
Article 70	Plant phytochemicals as epigenetic modulators: role in cancer chemoprevention.	54	49	75	23
Article 71	Problematic stabilizing films in petroleum emulsions: shear rheological response of viscoelastic asphaltene films and the effect on drop coalescence.	57	36	66	14
Article 72	Targeting of NAD metabolism in pancreatic cancer cells: potential novel therapy for pancreatic tumors.	54	18	65	13
Article 73	Myasthenia Gravis: paradox versus paradigm in autoimmunity.	56	124	58	6
Article 74	Mendelian randomization in health research: using appropriate genetic variants and avoiding biased estimates.	56	41	57	5
Article 75	Ag(x)@WO_3 core-shell nanostructure for LSP enhanced chemical sensors.	54	16	56	4
Article 76	Predictably selective (sp3)C-O bond formation through copper catalyzed dehydrogenative coupling: facile synthesis of dihydro-oxazinone derivatives.	54	31	64	13
Article 77	The origin of segmentation motor activity in the intestine.	52	15	66	16
Article 78	Phosphotungstic acid encapsulated in the mesocages of amine-functionalized metal-organic frameworks for catalytic oxidative desulfurization.	52	21	63	13
Article 79	Prefrontal cortical GABAergic dysfunction contributes to age-related working memory impairment.	53	34	51	2
Article 80	Directed evolution of an ultrastable carbonic anhydrase for highly efficient carbon capture from flue gas.	49	16	56	7

Article 81	Tumor necrosis factor alpha in mycobacterial infection.	52	15	52	3
Article 82	Epigenetic biomarkers in urological tumors: A systematic review.	49	232	51	2
Article 83	MicroRNA-145: a potent tumour suppressor that regulates multiple cellular pathways.	45	8	52	4
Article 84	Vitamin B_{12}-containing plant food sources for vegetarians.	40	27	51	4
Article 85	Complex interaction of dendritic connectivity and hierarchical patch size on biodiversity in river-like landscapes.	46	65	54	8
Article 86	The autophagy regulators Ambra1 and Beclin 1 are required for adult neurogenesis in the brain subventricular zone.	49	25	49	3
Article 87	MicroRNA-145 suppresses hepatocellular carcinoma by targeting IRS1 and its downstream Akt signaling.	45	12	55	10
Article 88	Molecular mechanisms of endothelial NO synthase uncoupling.	44	9	53	8
Article 89	Defect-free, size-tunable graphene for high-performance lithium ion battery.	48	21	52	7
Article 90	Social support predicts inflammation, pain, and depressive symptoms: longitudinal relationships among breast cancer survivors.	45	68	52	7
Article 91	Animal models of CNS disorders.	47	6	50	6
Article 92	The airway microbiome of intubated premature infants: characteristics and changes that predict the development of bronchopulmonary dysplasia.	45	41	49	5
Article 93	On the evolutionary origins of obesity: a new hypothesis.	44	61	65	22
Article 94	Predictive value of methicillin-resistant Staphylococcus aureus (MRSA) nasal swab PCR assay for MRSA pneumonia.	31	13	47	5
Article 95	Impact of postoperative non-steroidal anti-inflammatory drugs on adverse events after gastrointestinal surgery.	41	9	48	6
Article 96	DNA methylation markers for early detection of women's cancer: promise and challenges.	38	12	51	10
Article 97	Evaluation of predictions in the CASP10 model refinement category.	49	17	48	8
Article 98	SALL4, a novel marker for human gastric carcinogenesis and metastasis.	46	32	65	26

Article 99	Therapeutic applications of curcumin for patients with pancreatic cancer.	42	7	49	11
Article 100	Autoantibody biomarkers in childhood-acquired demyelinating syndromes: results from a national surveillance cohort	42	33	43	12
Total		7317	9795	8209	890
Avg		73.17	97.95	82.09	8.9
Max		179	642	200	40
Min		31	6	43	0

Appendix B: Ranking of the sampled articles based on the proposed direct citation weight

	citation #	# of citation mentions	mentions>1	# citation mentions>1	Positive	Proposed Metric (using classifications)
1	180	326	70	216	241	249.96
2	167	215	29	77	146	197.92
3	162	267	51	156	135	241.03
4	157	194	29	66	39	178.50
5	149	204	34	89	114	190.58
6	137	256	47	166	119	215.72
7	128	267	47	186	118	230.58
8	124	205	41	122	68	169.88
9	118	160	26	68	79	142.73
10	117	159	20	62	90	148.75
11	110	221	41	152	113	202.24
12	110	169	33	92	64	149.21
13	107	141	21	55	80	130.91
14	107	139	21	53	57	126.74
15	102	172	32	102	58	156.56
16	102	155	23	76	54	140.26
17	101	168	33	100	78	145.02
18	100	156	33	89	77	135.33
19	100	133	23	56	59	119.69
20	97	164	32	99	73	142.48
21	96	180	32	116	56	157.67
22	96	180	45	129	53	148.36
23	94	118	20	44	45	107.25
24	93	252	55	214	127	212.52

25	93	131	16	54	63	117.82
26	88	118	18	48	57	112.50
27	87	161	35	109	80	139.23
28	85	184	32	131	94	170.14
29	85	102	13	30	21	96.00
30	82	146	31	95	53	124.04
31	80	111	20	51	47	103.42
32	80	100	10	30	41	93.38
33	78	98	15	35	47	92.92
34	77	108	18	49	64	93.63
35	75	133	25	83	64	116.68
36	74	96	15	37	57	90.33
37	74	87	11	24	49	82.00
38	73	87	9	23	62	81.40
39	73	86	13	26	29	79.50
40	72	151	29	108	76	122.13
41	72	127	26	81	67	113.40
42	71	100	22	51	49	85.50
43	70	80	7	17	35	77.25
44	69	105	17	53	63	95.51
45	68	106	20	58	51	88.33
46	66	87	14	35	54	76.83
47	65	105	24	64	47	97.90
48	64	100	18	54	50	88.55
49	64	74	8	18	49	69.25
50	63	107	15	59	56	97.74
51	63	99	22	58	36	87.88
52	62	114	14	66	53	97.89

53	61	94	19	52	52	81.96
54	61	88	22	49	37	79.33
55	60	77	11	28	21	69.42
56	59	110	16	67	66	91.95
57	59	90	23	54	51	71.58
58	58	138	25	105	52	117.66
59	57	125	23	91	91	107.01
60	57	106	21	70	44	83.85
61	56	83	13	40	39	73.21
62	56	67	10	21	20	61.00
63	54	105	18	69	42	92.84
64	54	91	22	59	42	74.58
65	54	72	11	29	39	65.17
66	54	68	10	24	37	64.92
67	54	67	9	22	35	61.42
68	54	64	9	19	32	58.75
69	53	104	21	72	44	90.88
70	52	87	18	53	34	79.24
71	52	78	16	42	33	71.21
72	52	75	13	36	32	66.40
73	52	70	10	28	31	66.01
74	52	64	7	19	28	62.00
75	52	66	12	26	24	59.00
76	51	68	10	27	45	64.47
77	50	66	12	28	35	59.08
78	50	61	8	19	26	57.75
79	49	98	17	66	56	84.68
80	49	73	16	40	52	64.42

81	49	62	10	23	36	57.50
82	49	60	9	20	32	56.25
83	48	56	8	16	29	52.00
84	47	63	9	25	39	57.54
85	46	78	16	48	53	69.04
86	46	66	15	35	23	55.67
87	45	74	12	41	44	66.48
88	45	56	10	21	32	50.75
89	45	58	7	20	21	50.42
90	45	53	7	15	16	48.25
91	44	66	12	34	36	59.83
92	44	50	6	12	15	46.50
93	43	72	14	43	30	63.15
94	42	54	11	23	28	51.00
95	42	50	7	15	23	46.50
96	41	51	8	18	25	48.75
97	40	54	11	25	28	49.25
98	39	61	13	35	34	53.08
99	38	44	5	11	27	42.75
100	31	47	9	25	25	41.29

Appendix C: Rank Change with Number of Citations and the Proposed Metrics

	No of citations	Propose Metric from semantic similarity classes	Change in rank (Propose Metric from semantic similarity classes)
1	180	249.96	0
2	167	197.92	5
3	162	241.03	1
4	157	178.50	5
5	149	190.58	3
6	137	215.72	2
7	128	230.58	4
8	124	169.88	3
9	118	142.73	9
10	117	148.75	5
11	110	202.24	5
12	110	149.21	2
13	107	130.91	10
14	107	126.74	10
16	102	140.26	4
15	102	156.56	2
17	101	145.02	0
19	100	119.69	8
18	100	135.33	4
20	97	142.48	1
21	96	157.67	9
22	96	148.36	6
23	94	107.25	10
25	93	117.82	3
24	93	212.52	19
26	88	112.50	6
27	87	139.23	6
28	85	170.14	18
29	85	96.00	10
30	82	124.04	5
31	80	103.42	4
32	80	93.38	10
33	78	92.92	10
34	77	93.63	7
35	75	116.68	5
36	74	90.33	11
37	74	82.00	17

38	73	81.40	18
39	73	79.50	18
41	72	113.40	10
40	72	122.13	14
42	71	85.50	9
43	70	77.25	17
44	69	95.51	4
45	68	88.33	4
46	66	76.83	15
47	65	97.90	11
48	64	88.55	0
49	64	69.25	18
50	63	97.74	12
51	63	87.88	1
52	62	97.89	15
54	61	79.33	4
53	61	81.96	2
55	60	69.42	11
57	59	71.58	7
56	59	91.95	11
58	58	117.66	29
59	57	107.01	25
60	57	83.85	7
61	56	73.21	2
62	56	61.00	17
64	54	74.58	2
66	54	64.92	7
65	54	65.17	7
67	54	61.42	11
63	54	92.84	19
68	54	58.75	15
69	53	90.88	23
73	52	66.01	2
74	52	62.00	3
70	52	79.24	11
71	52	71.21	6
72	52	66.40	2
75	52	59.00	7
76	51	64.47	2
78	50	57.75	6
77	50	59.08	4
82	49	56.25	5

79	49	84.68	27
80	49	64.42	5
81	49	57.50	5
83	48	52.00	7
84	47	57.54	1
85	46	69.04	17
86	46	55.67	2
90	45	48.25	6
89	45	50.42	4
87	45	66.48	18
88	45	50.75	4
92	44	46.50	5
91	44	59.83	11
93	43	63.15	17
94	42	51.00	3
95	42	46.50	3
96	41	48.75	1
97	40	49.25	3
98	39	53.08	9
99	38	42.75	0
100	31	41.29	0

Appendix D: Change in rank by the number of Citation mentions and the proposed Citation weights

	# of citation mentions	Propose Metric from semantic similarity classes	Change in rank (Propose Metric from semantic similarity classes)
1	326	249.96	0
2	267	241.03	0
3	267	230.58	0
4	256	215.72	0
5	252	212.52	0
6	221	202.24	0
7	215	197.92	0
8	205	169.88	3
9	204	190.58	1
10	194	178.50	1

11	184	170.14	1
12	180	157.67	0
13	180	148.36	3
14	172	156.56	1
15	169	149.21	1
16	168	145.02	1
17	164	142.48	2
18	161	139.23	3
19	160	142.73	1
20	159	148.75	5
21	156	135.33	1
22	155	140.26	2
23	151	122.13	3
24	146	124.04	1
25	141	130.91	2
26	139	126.74	2
27	138	117.66	2
28	133	119.69	1
29	133	116.68	1
30	131	117.82	2
31	127	113.40	0
32	125	107.01	2
33	118	112.50	2
34	118	107.25	0
35	114	97.89	2
36	111	103.42	1
37	110	91.95	8
38	108	93.63	3
39	107	97.74	1
40	106	88.33	9
41	106	83.85	12
42	105	97.90	7
43	105	95.51	2
44	105	92.84	0
45	104	90.88	1
46	102	96.00	7
47	100	93.38	5
48	100	88.55	1
49	100	85.50	3
50	99	87.88	0
51	98	92.92	8
52	98	84.68	0

53	96	90.33	6
54	94	81.96	1
55	91	74.58	7
56	90	71.58	8
57	88	79.33	1
58	87	82.00	4
59	87	81.40	3
60	87	79.24	2
61	87	76.83	1
62	86	79.50	5
63	83	73.21	0
64	80	77.25	4
65	78	71.21	0
66	78	69.04	2
67	77	69.42	1
68	75	66.40	2
69	74	69.25	2
70	74	66.48	1
71	73	64.42	4
72	72	65.17	0
73	72	63.15	3
74	70	66.01	3
75	68	64.92	2
76	68	64.47	2
77	67	61.42	0
78	67	61.00	2
79	66	59.83	2
80	66	59.08	1
81	66	59.00	3
82	66	55.67	7
83	64	62.00	7
84	64	58.75	0
85	63	57.54	0
86	62	57.50	0
87	61	57.75	3
88	61	53.08	1
89	60	56.25	2
90	58	50.42	3
91	56	52.00	1
92	56	50.75	0
93	54	51.00	2
94	54	49.25	0

95	53	48.25	1
96	51	48.75	1
97	50	46.50	0
98	50	46.50	0
99	47	41.29	1
100	44	42.75	1

Appendix E: Change in Rank by the number of Multiple Citation mentions and the proposed citation weight

	No of mentions>1	Proposed Metric (using classifications)	Change in rank (Propose Metric from semantic similarity classes)
1	70	249.96	0
2	55	212.52	3
3	51	241.03	1
5	47	230.58	2
4	47	215.72	0
6	45	148.36	10
8	41	202.24	2
7	41	169.88	4
9	35	139.23	12
10	34	190.58	2
12	33	145.02	5
11	33	149.21	3
13	33	135.33	9
17	32	170.14	7
16	32	157.67	4
15	32	142.48	4
14	32	156.56	1
18	31	124.04	7
21	29	122.13	5
19	29	197.92	12
20	29	178.50	11
23	26	113.40	8
22	26	142.73	4
25	25	117.66	4
24	25	116.68	6
26	24	97.90	10
27	23	140.26	7
30	23	107.01	4
28	23	119.69	1
29	23	71.58	35
34	22	74.58	28
32	22	87.88	18
33	22	79.33	25
31	22	85.50	20

38	21	90.88	8
37	21	83.85	16
35	21	130.91	12
36	21	126.74	12
39	20	148.75	24
41	20	103.42	6
42	20	88.33	7
40	20	107.25	7
43	19	81.96	12
46	18	88.55	2
44	18	112.50	12
45	18	93.63	4
48	18	79.24	11
47	18	92.84	3
49	17	95.51	9
50	17	84.68	2
51	16	117.82	23
55	16	69.04	13
52	16	91.95	7
53	16	71.21	12
54	16	64.42	21
58	15	97.74	20
57	15	90.33	10
56	15	92.92	13
59	15	55.67	29
61	14	97.89	24
62	14	63.15	14
60	14	76.83	1
65	13	73.21	2
67	13	53.08	22
63	13	96.00	24
66	13	66.40	4
64	13	79.50	7
68	12	59.00	14
69	12	59.08	12
70	12	66.48	1
71	12	59.83	9
74	11	65.17	2
73	11	69.42	7
76	11	49.25	18
75	11	51.00	16
72	11	82.00	18

80	10	66.01	9
79	10	64.92	6
77	10	93.38	35
81	10	64.47	7
82	10	57.50	4
78	10	61.00	1
83	10	50.75	9
87	9	56.25	0
88	9	57.54	3
85	9	61.42	7
89	9	41.29	11
84	9	81.40	28
86	9	58.75	3
91	8	57.75	7
93	8	48.75	2
90	8	69.25	23
92	8	52.00	2
97	7	48.25	1
95	7	62.00	18
96	7	50.42	3
94	7	77.25	34
98	7	46.50	1
99	6	46.50	1
100	5	42.75	1

Appendix F: Sampled articles ranked by the sum of multiple citations

	Sum of multiple citation mentions	Propose Metric from semantic similarity classes	Change in rank (Propose Metric from semantic similarity classes)
1	216	249.96	0
2	214	212.52	3
3	186	230.58	0
4	166	215.72	0
5	156	241.03	3
6	152	202.24	0
7	131	170.14	3
8	129	148.36	8
9	122	169.88	2
10	116	157.67	2
11	109	139.23	10
12	108	122.13	14
13	105	117.66	16
14	102	156.56	1
15	100	145.02	2
16	99	142.48	3
17	95	124.04	8
18	92	149.21	4
19	91	107.01	15
20	89	190.58	12
21	89	135.33	1
22	83	116.68	8
23	81	113.40	8
24	77	197.92	17
25	76	140.26	5
26	72	90.88	20
27	70	83.85	26
28	69	92.84	16
29	68	142.73	11
30	67	91.95	15
32	66	97.89	5
31	66	178.50	22
33	66	84.68	19
34	64	97.90	2
35	62	148.75	20
37	59	74.58	25
36	59	97.74	2

39	58	87.88	11
38	58	88.33	11
40	56	119.69	13
41	55	130.91	18
42	54	117.82	14
44	54	71.58	20
43	54	88.55	5
46	53	95.51	6
45	53	126.74	21
47	53	79.24	12
48	52	81.96	7
49	51	103.42	14
50	51	85.50	1
52	49	79.33	6
51	49	93.63	10
54	48	69.04	14
53	48	112.50	21
55	44	107.25	22
56	43	63.15	20
57	42	71.21	8
58	41	66.48	11
59	40	73.21	4
60	40	64.42	15
61	37	90.33	14
62	36	66.40	8
63	35	92.92	20
66	35	53.08	23
64	35	76.83	3
65	35	55.67	23
67	34	59.83	13
69	30	93.38	27
68	30	96.00	29
70	29	65.17	2
72	28	66.01	1
71	28	69.42	5
73	28	59.08	8
74	27	64.47	0
76	26	59.00	6
75	26	79.50	18
77	25	57.54	8
79	25	41.29	21
78	25	49.25	16

81	24	64.92	8
80	24	82.00	26
82	23	81.40	26
84	23	51.00	7
83	23	57.50	3
85	22	61.42	7
86	21	61.00	7
87	21	50.75	5
88	20	56.25	1
89	20	50.42	4
91	19	62.00	14
92	19	57.75	8
90	19	58.75	7
94	18	48.75	1
93	18	69.25	26
95	17	77.25	35
96	16	52.00	6
97	15	48.25	1
98	15	46.50	1
99	12	46.50	1
100	11	42.75	1

Appendix G: Ranking by the number of Positive Citation Sentiments

	Positive citation sentiment	Propose Metric from semantic similarity classes	Change in rank (Propose Metric from semantic similarity classes)
1	241	249.96	0
2	146	197.92	5
3	135	241.03	1
4	127	212.52	1
5	119	215.72	1
6	118	230.58	3
7	114	190.58	1
8	113	202.24	2
9	94	170.14	1
10	91	107.01	24
11	90	148.75	4
13	80	139.23	8
12	80	130.91	11
14	79	142.73	4
15	78	145.02	2
16	77	135.33	6
17	76	122.13	9
18	73	142.48	1
19	68	169.88	8
20	67	113.40	11
21	66	91.95	24
24	64	116.68	6
22	64	149.21	8
23	64	93.63	18
25	63	117.82	3
26	63	95.51	14
27	62	81.40	29
28	59	119.69	1
29	58	156.56	16
31	57	112.50	1
30	57	126.74	6
32	57	90.33	15
33	56	157.67	21
34	56	97.74	4
35	56	84.68	17
36	54	140.26	16

37	54	76.83	24
38	53	148.36	22
39	53	124.04	14
40	53	97.89	3
41	53	69.04	27
43	52	117.66	14
42	52	81.96	13
44	52	64.42	31
46	51	71.58	18
45	51	88.33	4
47	50	88.55	1
49	49	85.50	2
48	49	82.00	6
50	49	69.25	17
53	47	97.90	17
51	47	103.42	16
52	47	92.92	9
54	45	107.25	21
55	45	64.47	19
57	44	90.88	11
56	44	83.85	3
58	44	66.48	11
60	42	74.58	2
59	42	92.84	15
61	41	93.38	19
62	39	178.50	53
63	39	73.21	0
64	39	65.17	8
65	39	57.54	20
67	37	64.92	6
66	37	79.33	8
68	36	87.88	18
69	36	57.50	17
70	36	59.83	10
72	35	61.42	6
73	35	59.08	8
71	35	77.25	11
74	34	79.24	15
75	34	53.08	14
76	33	71.21	11
79	32	56.25	8
78	32	66.40	8

77	32	58.75	6
80	32	50.75	12
81	31	66.01	10
82	30	63.15	6
83	29	79.50	26
84	29	52.00	6
85	28	62.00	8
87	28	49.25	7
86	28	51.00	5
88	27	42.75	11
89	26	57.75	5
91	25	41.29	9
90	25	48.75	5
92	24	59.00	10
93	23	55.67	5
94	23	46.50	3
95	21	96.00	56
96	21	69.42	30
97	21	50.42	4
98	20	61.00	19
99	16	48.25	3
100	15	46.50	2

Curriculum Vitae

Name: Toluwase Asubiaro

Post-secondary Education and Degrees:

University of Ado-Ekiti,
Ado-Ekiti, Ekiti State, Nigeria
1999-2006 B. Sc. (Hons) Mathematics

University of Ibadan,
Ibadan, Oyo State, Nigeria
2009-2011 M. Inf. Sc.

The University of Western Ontario
London, Ontario, Canada
2016-2021 Ph.D.

Honours and Awards

Eugene Garfield Research Fellowship by Medical Library Association
2020-2021

Ontario Graduate Scholarships and the Queen Elizabeth II Graduate Scholarships in Science and Technology (OGS/QEII-GSST)
2019-2020

Western Graduate Research Scholarship
Doctoral Scholarship
2016-2020

Related Work Experience

Teaching Assistant
Faculty of Information and Media Studies
University of Western Ontario
2016-2020

Research Assistant
Language and Information Technology Research Laboratory
Faculty of Information and Media Studies
University of Western Ontario
Canada
2016-2020

Research Assistant
Unbundling Big deals Research Laboratory
Faculty of Information and Media Studies

University of Western Ontario
Canada
2020-2021

Limited Duty Instructor
LIS 9002-Information Organization, Curation and Access
Faculty of Information and Media Studies
University of Western Ontario
Canada
2020 Fall Term

Limited Duty Instructor
LIS 9701-Information Retrieval: Research and Practice
Faculty of Information and Media Studies
University of Western Ontario
Canada
2021 Winter Term

Publications:

Asubiaro, T.V., & Okonkwo-Elueze, I. (unpublished) Role of librarians from Sub-Saharan Africa in evidence-based Biomedical Research: An Analysis of Meta-Analysis and Systematic Review articles in MEDLINE.

Asubiaro, T., Mongeon, P. & Simard, M. (unpublished) Big Deal 2.0? Exploring the impact of Hybrid/Open Access Journals' Articles Processing Charges on Information Access Cost.

Asubiaro, T.V. & Shaik, H. (submitted) Evaluating Sub-Saharan Africa's COVID-19 Research Contribution: A Preliminary bibliometric Analysis. *Scientific African*

Asubiaro, T. (2021). Evaluating the Availability of Resources, Research Hubs and Financial Supports for Nigerian Languages Natural Language Processing Research. *Canadian Journal of Library and Information Science*

Asubiaro, T., Badmus, O., Ikenyei, U., Popoola, B., & Igwe, E. (2021). Exploring Sub-Saharan Africa's communication of COVID-19-Related Health Information on Social Media. *Libri-International Journal of Libraries and Information Studies*. 71 (2)

Rubin, V., Burkell, J., Cornwell, S., **Asubiaro, T**., Chen, Y., Potts, D., & Brogly, C. (2020,). AI Opaqueness: What Makes AI Systems More Transparent? *Proceedings of the Annual Conference of CAIS / Actes Du congrès Annuel De l'ACSI*. https://doi.org/10.29173/cais1139

Asubiaro T.V. and Badmus O.M. (2020). Collaboration clusters, interdisciplinarity, scope and subject classifications of Library and Information Science Research from Africa: An analysis of Web of Science Publications from 1996 to 2015. *Journal of Librarianship and Information Science.* 52(4) Pp. 1169-1185 https://doi.org/10.1177/0961000620907958

Asubiaro, T. V. (2019). How Collaboration Type, Publication Place, Funding and Author's role affect Citations Received by Publications from Africa: A Bibliometric study of LIS research from 1996 to 2015. *Scientometrics* 120(3) pp 1261–1287. https://doi.org/10.1007/s11192-019-03157-1

Rubin, V.L.; Brogly, C.; Conroy, N.; Chen, Y.; Cornwell, S. and **Asubiaro, T. V.** (2019). A News Verification Browser for the Detection of Clickbait, Satire and Falsified News. *Journal of Open Source Software* 4(35) 1208. https://doi.org/10.21105/joss.01208

Asubiaro, Toluwase; Adegbola, Tunde; Mercer, Robert and Ajiferuke, Isola (2018). A Word-Level Language Identification Strategy for Resource-Scarce Languages. *Proceedings of the Association for Information Science and Technology*, 55(1), 19-28. https://doi.org/10.1002/pra2.2018.14505501004

Asubiaro, Toluwase; Rubin, Victoria (2018). Comparing Features of Fabricated and Legitimate Political News in Digital Environments (2016-2017). *Proceedings of the Association for Information Science and Technology*, 55(1), 747-750. https://doi.org/10.1002/pra2.2018.14505501100

Asubiaro, Toluwase (2018). Research Collaboration Landscape of the University of Ibadan Biomedical Authors between 2006 and 2015. *African Journal of Library, Archives and Information Science*, Vol. 28, No 1, (April 2018) 17-31

Asubiaro, Toluwase (2017). An Assessment of the Cyber Presence of Academic Libraries in Nigeria. *African Journal of Library, Archives and Information Science*, Vol. 27, No. 1, (April 2017) 65-76

Asubiaro, Toluwase (2014). Effects of Diacritics on Web Search Engines' Performance for Retrieval of Yoruba Documents. *Journal of Library and Information Studies* 12:1 (June 2014) pp 1-18. http://doi.org/10.6182/jlis.2014.12(1).001

Asubiaro, Toluwase (2013). Entropy-Based Generic Stopwords List for Yoruba Texts. *International Journal of Computer and Information Technology.* Vol. 2(5), Pp. 1065-1068