

**PREDICTING THE CITATION COUNTS OF
RESEARCH PAPERS USING NEURAL NETWORKS**

I.M.D.Dayarathna

199314V

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

March 2021

DECLARATION

“I declare that this is my own work and this thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).”

UOM Verified Signature

Signature:

Date: 29/05/2021

The supervisor/s should certify the thesis with the following declaration.

The above candidate has carried out research for the Masters thesis under my supervision.

Name of the supervisor: Dr. Charith Chitraranjan

UOM Verified Signature

Signature of the supervisor:

Date: 29/05/2021

ABSTRACT

A widely accepted criterion used to measure the scientific impact of a research paper is the citation count. However, for a newly published paper this metric does not become available for several years after the date of publication. Yet, many parties including fellow scholars, research institutes and funding bodies find it important to be able to identify early on, the scientific papers with a higher potential to make a bigger impact. Predicting the future citation counts is an effective solution to overcome this limitation.

However, predicting the future citation count of a scientific paper is a challenging task, particularly due to the highly dynamic nature in the citation accumulation process. Hence this remains an active area of research. A majority of the prior studies that predict future citation counts using features available at the time of publication of a research paper make use of classical machine learning techniques. In this study, the author demonstrates through experiments how artificial neural network models can outperform best performing classical machine learning models discussed in prior studies.

One notable limitation of current approaches to this research problem is that many approaches treat the citation networks as unweighted graphs. In this work, the author demonstrates how treating the citation relationships as weighted relationships could help improve performance of the models. For this, the author introduces a novel feature named Weighted Average Neighboring Citation Score, a value computed by treating the citation network as a weighted graph, and demonstrates through multiple experiments that the newly introduced feature helps improve the performance of multiple models. Moreover, the author experiments with different edge weighting schemes and demonstrates how factoring both the recency of a citation and frequency with which a source has been cited when determining the edge weights help improve the performance of the models.

Keywords: Citation Count Prediction, Neural Networks, Weighted Citation Networks

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my supervisor, Dr. Charith Chitraranjan, for all the invaluable guidance provided to make this project a success.

I would also like to express my sincere gratitude to my family and friends for encouraging me on numerous occasions during this endeavor.

TABLE OF CONTENTS

DECLARATION.....	i
ABSTRACT.....	ii
ACKNOWLEDGMENTS.....	iii
LIST OF FIGURES.....	vi
LIST OF TABLES.....	vii
LIST OF ABBREVIATIONS.....	viii
LIST OF APPENDICES.....	viii
1 INTRODUCTION.....	1
1.1 The Need to Evaluate the Scientific Impact of Research.....	1
1.2 Evaluating the Scientific Research through Peer-Reviews.....	3
1.3 Using the Citation Count as a Proxy to Measure Scientific Impact.....	3
1.4 The Need to Predict the Citation Counts Upfront.....	4
1.5 Challenging Nature of the Problem of Predicting Citation Counts.....	5
1.6 Existing Approaches to Predict the Citation Counts and Gaps in Current Research.....	6
1.7 Research Questions.....	8
1.8 Objectives of the Research.....	8
1.9 Contributions of the Research.....	8
2 LITERATURE REVIEW.....	10
2.1 Background.....	10
2.2 Classical Machine Learning Based Approaches.....	10
2.2.1 A Brief Summary of Commonly Used Features.....	11
2.2.2 A Summary of the Key Papers that use Classical Machine Learning Techniques to Predict the Future Citation Counts.....	13
2.2.3 Classical Machine Learning Models that use AltMetrics to Predict the Citation Count.....	17
2.3 Approaches using Artificial Neural Networks.....	18
2.4 Graph Based Approaches.....	21
2.5 Time Series Analysis Based Approaches.....	22
3 METHODOLOGY.....	24
3.1 Research Questions.....	24

3.2	Creating the Dataset.....	24
3.3	Pre-Processing the Dataset.....	27
3.4	Feature Engineering.....	28
3.4.1	Introducing a Novel Feature - Weighted Average Neighboring Citation Score	29
3.4.2	Author level and Venue Level Features	32
3.4.3	Additional Processing of Data and Data Transformations	33
3.5	Visualizing the Dataset	33
3.6	Research Methods.....	38
3.7	Model Descriptions.....	39
3.7.1	Multilayer Perceptron Neural Network Model	39
3.7.2	Graph Neural Network Models	40
3.7.3	Baseline Models	40
3.8	Experimental Setup.....	40
3.9	Evaluation Metric	41
4	RESULTS AND DISCUSSION.....	42
4.1	Answering the Research Question 1.....	42
4.2	Answering the Research Question 2.....	48
4.3	Discussion	50
	REFERENCES	52
	APPENDIX A: Excerpts from the Source Code Used To Extract the Dataset.....	59
	APPENDIX B: Excerpts from the Source Code Used To Derive Key Features	60
	APPENDIX C: Excerpts from the Code Used To Build the Multilayer Perceptron Neural Network Models.....	61
	APPENDIX D: Excerpts from the Code Used to Build the GraphSAGE Models.....	64
	APPENDIX E: Excerpts from the Code Used to Build the TAGCN Models with Edge Weights	66
	APPENDIX F: Excerpts from the Code Used to Fit the Linear Regression Models	67
	APPENDIX G: Excerpts from the Code Used to Fit the SVR Models	68
	APPENDIX H: Excerpts from the Code Used to Test the Statistical Significance of the Results of the Multilayer Perceptron Neural Network Model.....	69

LIST OF FIGURES

	Page
Figure 1.1 Number of research publications in all fields over the time	1
Figure 1.2 Number of published authors in all fields over the time	2
Figure 3.1 Number of publications in the field of AI over time	25
Figure 3.2 A sample request sent to the API	26
Figure 3.3 A sample of citation contexts in the paper 2153738822	26
Figure 3.4 Distribution of the citation counts in full dataset	33
Figure 3.5 Distribution of the citation counts of a filtered sample where the maximum citation count is 500	34
Figure 3.6 Distribution of the papers by publication year	34
Figure 3.7 Distribution of the number of authors per paper	34
Figure 3.8 Distribution of total citation counts of the first authors	35
Figure 3.9 Distribution of total publication counts of the first authors	35
Figure 3.10 Distribution of the h-index vales of the first authors	35
Figure 3.11 Distribution of the total citation counts of the venues	36
Figure 3.12 Distribution of the total publication counts of the venues	36
Figure 3.13 How citation count varies with the h-index of the first author	36
Figure 3.14 How citation count varies with the venue rank	37
Figure 3.15 How citation count varies with the paper age	37
Figure 3.16 How citation count varies with the number of authors	37
Figure 3.17 How citation count varies with the frequency weighted average neighboring citation score	38
Figure 3.18 How citation count varies with the frequency and time weighted average neighboring citation score	38
Figure 4.1 Performance of the models when predicting the citation count	44

after 5 years

Figure 4.2 Performance of the models when predicting the citation count after 10 years	45
Figure 4.3 Performance of the models when predicting the citation count after 15 years	45
Figure 4.4 Performance of the TAGCN model before and after assigning the citation edge weights	46
Figure 4.5 LR Model: comparing the actual vs. predicted log citation counts after 5 years	46
Figure 4.6 SVR Model: comparing the actual vs. predicted log citation counts after 5 years	47
Figure 4.7 NN Model: comparing the actual vs. predicted log citation counts after 5 years	47
Figure 4.8 GraphSAGE Model: comparing the actual vs. predicted log citation counts after 5 years	47

LIST OF TABLES

	Page
Table 3.1 An overview of the training and test datasets used	28
Table 4.1 Performance of the models when predicting the citation count after 5 years	43
Table 4.2 Performance of the models when predicting the citation count after 10 years	43
Table 4.3 Performance of the models when predicting the citation count after 15 years	44
Table 4.4 Performance of the TAGCN model before and after assigning the citation edge weights	44

LIST OF ABBREVIATIONS

Abbreviation	Description
LR	Linear Regression
NN	Neural Network
SVR	Support Vector Regression
MLP	Multilayer Perceptron
GNN	Graph Neural Network

LIST OF APPENDICES

Appendix	Description	Page
Appendix A	Excerpts from the Source Code Used To Extract the Dataset	59
Appendix B	Excerpts from the Source Code Used To Derive Key Features	60
Appendix C	Excerpts from the Code Used To Build the Multilayer Perceptron Neural Network Models	61
Appendix D	Excerpts from the Code Used To Build the GraphSAGE Models	64
Appendix E	Excerpts from the Code Used To Build the TAGCN Models with Edge Weights	66
Appendix F	Excerpts from the Code Used to Fit the Linear Regression Models	67
Appendix G	Excerpts from the Code Used to Fit the SVR Models	68
Appendix H	Excerpts from the Code Used to Test the Statistical Significance of the Results of the Multilayer Perceptron Neural Network Model	69