University of Technology Sydney
Faculty of Engineering and Information Technology
School of Electrical and Data Engineering

Submodular Approaches for Citation Recommendation

Thanh Binh Kieu

A Thesis Submitted
in Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

under the supervision of

Prof. Massimo Piccardi
Dr. Diep N. Nguyen
Dr. Hieu Xuan Phan
Dr. Son Bao Pham

Sydney, Australia

March 2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Thanh Binh Kieu, declare that this thesis is submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of the requirements for a degree at any other academic institution except as fully acknowledged within the text. This thesis is the result of a Collaborative Doctoral Research Degree program with Vietnam National University Hanoi.

This research is supported by the Australian Government Research Training Program.

Signature:

Date: 07/03/2022

ABSTRACT

Submodular Approaches for

Citation Recommendation

by

Thanh Binh Kieu

With the rapid growth of the scientific literature, manually selecting appropriate citations for a paper is becoming increasingly challenging and time-consuming. Automated citation recommendation can help ease this challenge by suggesting the most appropriate citations for a query document, e.g., a thesis draft. While several approaches for automated citation recommendation have been proposed in the recent years, effective improvements in content-based citation recommendation are still elusive to a large extent. In this thesis, we aim to find a novel approach for recommending global citations for an academic paper draft based on deep representation learning and submodular inference. The current state-of-the-art systems for this task are based on deep learning and graph representations and have already achieved impressive results. However, their results are only based on the ranking of matching scores and do not fully answer the question: "is the recommended list of references adequate?". In this thesis, we aim to provide an answer to this question by applying citation selection methods instead of matching score rankings only. For this, we rely on submodular inference in combination with a deep sequential representation of the query and the candidate references. The results we have obtained show that the proposed approach has been able to outperform a range of existing, relevant approaches over a number of citation datasets and evaluation measures.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors Prof. Massimo Piccardi, Dr. Diep N. Nguyen, Dr. Xuan-Hieu Phan, Dr. Bao-Son Pham for all their guidance, encouragement, and enduring patience. Prof. Massimo has taught me a lot on both the scientific and personal sides of the research journey. Their guidance and support go far beyond this thesis, and I have been greatly fortunate to be supervised by them.

Furthermore, many thanks to my friend and colleague Iñigo Jauregi Unanue and all the staff from the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology, University of Technology Sydney (UTS). Special thanks to the Joint Research Center between UTS and VNU for giving me the rare opportunity to study in both prestigious universities in Vietnam and Australia throughout my whole PhD study.

Last but most importantly, my deepest gratitude, love and respect go to my whole families who have been supporting and encouraging me in every possible way as they had always done. Their love, sacrifice, and patience will be the constant source of motivation throughout my life.

Contents

Certificate of Original Authorship ii

Abstract iii

Acknowledgments iv

Table of Contents v

List of Publications ix

List of Figures x

List of Tables xi

Abbreviation xiii

1 Introduction 1

1.1 Motivation 1

1.2 Objectives 4

1.3 Research questions 5

1.4 Contributions 5

1.5 Thesis Structure 6

2 Literature Review 9

2.1 Task and Approaches 9

2.1.1 Task Definition 9

2.1.2 Common Approaches and Techniques 12

2.1.3 Collaborative Filtering Approach 14

2.1.4 Content-Based Filtering Approach 16

2.1.5 Neural Network Techniques 18

2.1.6 Graph-Based Techniques 20

2.2 Submodular Inference 21
2.2.1 Definitions 22
2.2.2 Submodular Functions 24
2.3 Text Representations 26
2.3.1 Classic Techniques 26
2.3.2 Context-Independent Word Embeddings 30
2.3.3 Context-Aware Word Embeddings 35
2.3.4 Sentence Embeddings 45
2.3.5 Document Embeddings 46
2.4 Datasets and Metrics 47
2.4.1 Evaluation Metrics 47
2.4.2 Scholarly Datasets 53

3 Citation Recommendation by Submodular Inference 56
3.1 Introduction 56
3.2 Related Work 58
3.3 Submodularity Background 59
3.3.1 Definitions 59
3.3.2 Its Applications 61
3.4 Proposed Submodular Functions 63
3.4.1 Non-monotone Submodular Functions 63
3.4.2 Monotone Submodular Functions 64
3.5 Experiments 65
3.5.1 Corpus 65
3.5.2 Evaluation Metrics 66
3.5.3 Experimental Settings 67
3.5.4 Parameter Tuning 70
3.5.5 Performance Comparison 71

3.6 Conclusion 71

4 Citation Recommendation Based on Deep Representations 73

4.1 Introduction 73
4.2 Citation Recommendation Background 76
4.2.1 The Citation Recommendation Task 76
4.2.2 Document Scoring 77
4.2.3 Document Selection 78
4.3 The Proposed Approach: Learning Document Scoring 80
4.3.1 The Deep Textual Representation 80
4.3.2 The Network Architecture 81
4.3.3 The Training Approach 82
4.4 Experiments 84
4.4.1 Dataset 84
4.4.2 Compared Approaches 85
4.4.3 Evaluation Metrics 86
4.4.4 Main Results 87
4.4.5 Discussion 87
4.5 Conclusion 90

5 NeuSub: A Neural Submodular Approach for Citation Recommendation 92

5.1 Introduction 92
5.2 Related Work 95
5.2.1 Neural Network Approaches 97
5.2.2 Submodular Approaches 98
5.3 Problem Formulation 99

5.4 Methodology 101
5.4.1 Submodular Inference 102
5.4.2 Document Representation and Similarity 103
5.4.3 Submodularity-Oriented Training Objective 105
5.5 Experiments and Results 109
5.5.1 Datasets 109
5.5.2 Evaluation Metrics 110
5.5.3 Compared Approaches 112
5.5.4 Results 114
5.5.5 Comparison with graph embedding approaches 119
5.6 Conclusion 120

6 Conclusions and Future Research 122
6.1 Conclusions 122
6.2 Future Research 123

Bibliography 125

List of Publications

Following is the list of the refereed international journal and conference papers produced during my PhD research that have been published or are currently under review:

1. **Thanh-Binh Kieu**, Xuan-Hieu Phan, Son Bao Pham, Massimo Piccardi, "A Submodular Approach for Reference Recommendation", 16th International Conference of the Pacific Association for Computational Linguistics (PACLING), Hanoi, Vietnam, October 11–13, 2019.

2. **Thanh-Binh Kieu**, Inigo Jauregi Unanue, Son Bao Pham, Xuan-Hieu Phan, Massimo Piccardi, "Learning Neural Textual Representations for Citation Recommendation", 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, January 10-15, 2021.

3. **Thanh-Binh Kieu**, Inigo Jauregi Unanue, Son Bao Pham, Xuan-Hieu Phan, Massimo Piccardi, "NeuSub: A Neural Submodular Approach for Citation Recommendation", in IEEE Access, vol. 9, pp. 148459-148468, 2021.

List of Figures

1.1 An example of local citation recommendation. [Huang *et al.*, 2015] . . 3
1.2 Thesis Structure 7

2.1 The citation recommendation task: a query selects K citations from a corpus of candidates organized as a citation graph. 11
2.2 A content-based system for citation recommendation. 17
2.3 The NNLM structure [Bengio *et al.*, 2003]. 32
2.4 The CBOW and the skip-gram architectures [Mikolov *et al.*, 2013a]. . 33
2.5 Vietnamese-English translation with the Transformer. 37
2.6 An expanded encoder layer in the Vietnamese-English translation. . . 38
2.7 An example of Permutation Language Modeling [Yang *et al.*, 2019]. . 45
2.8 The SBERT architecture for classification (left) and regression (right) [Reimers and Gurevych, 2019]. 46

4.1 Left: document scoring with a Siamese network with a cosine similarity function; d_q: query document; d_i: candidate document. Right: an illustration of document selection with $K = 4$ 85

5.1 Overview of the proposed NeuSub training approach. 108
5.2 Overview of the proposed NeuSub inference. 108

List of Tables

1.1 Differences between global and local citation recommendation. 3

2.1 Summary of citation recommendation approaches. 13
2.2 Differences of content-based vs collaborative-based approaches. 14
2.3 Evaluation metric of reviewed papers. 49

3.1 Main statistics for the test set. 65
3.2 SubRef-QFRv1 by Lambda 68
3.3 SubRef-QFRv2 by Lambda 69
3.4 SubRef-QAIv1 by Lambda 69
3.5 SubRef-QAIv2 by Lambda 70
3.6 Performance comparison 70

4.1 Hyperparameters used for training the proposed approach 86
4.2 Results on the test set 88
4.3 Results for the Siamese network on the validation set with different selections of the training examples 89
4.4 Results for the triplet network on the validation set with different selections of the training examples 89
4.5 Comparison of different submodular functions (BM25 as the similarity score) 90

5.1 Main notations used in this chapter (approximately in order of appearance). 104
5.2 Main statistics of the citation datasets. 110
5.3 Minimum, maximum and average number of citations per document in the citation datasets. 110
5.4 Main hyperparameters used for training the proposed approach (NeuSub). 114
5.5 Results on the AAN test set. 115
5.6 Results on the DBLP test set. 115
5.7 Results on the PubMed test set. 116
5.8 An example from the AAN dataset: top: query (paper ID and title); two leftmost columns: ground-truth reference list of the query (paper ID and title); two rightmost columns: Citeomatic's and NeuSub's true predictions. 118
5.9 Comparison of attri2vec and NeuSub on the AAN, DBLP and PubMed test sets. 121

Abbreviation

ACL	Association for Computational Linguistics
AAN	ACL Anthology Network
AP	Average Precision
BERT	Bidirectional Encoder Representations from Transformer
BOW	Bag of Words
CBF	Content-Based Filtering
CBOW	Continuous Bag of Words
CF	Collaborative Filtering
DNN	Deep Neural Network
DL	Deep Learning
ELMo	Embeddings from Language Model
GloVe	Global Vector
GPT	Generative Pre-trained Transformer
IR	Information Retrieval
kNN	k-Nearest Neighbor
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
LSTM	Long Short-Term Memory
MAP	Mean Average Precision
ML	Machine Learning
MRR	Mean Reciprocal Rank
nDCG	normalized Discounted Cumulative Gain

NNLM	Neural Network Language Model
NLP	Natural Languge Processing
RNN	Recurrent Neural Network
SBERT	Sentence BERT
SVD	Singular Value Decomposition