



CROSS-DOMAIN LEARNING FOR NETWORK REPRESENTATIONS

Shan Xue

Faculty of Engineering and Information Technology

University of Technology Sydney

A thesis submitted for the Degree of

Doctor of Philosophy

February 2019

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

This thesis is the result of a research candidate conducted jointly with Shanghai University as part of a collaborative Doctoral degree. 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 except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

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

Production Note:

Signature removed
prior to publication.

Shan Xue

February 2019

Acknowledgements

On having completed this thesis, I would like to express my sincere gratitude to my principal supervisor Distinguished Professor Jie Lu. Without her patience and encouragement, I would not have been able to complete this PhD program. Professor Lu led me in a new academic research direction, giving me as much guidance as possible and trusted me to pursue my own research interests. She taught me how to think and study independently, and how to solve difficult scientific problems in a flexible and rigorous way. She instilled me with confidence and motivated me when I felt lost or afraid about the future. Her wisdom, decisiveness and sharp insights inspired me to keep going regardless of any difficulty I encountered. The help that Professor Lu gave me will continue to benefit me throughout my life.

I would also to express my thankfulness to my co-supervisor A./Professor Guangquan Zhang, who has been like a father to me throughout the duration of my four-year PhD candidature. I always knew that, in him, I had someone to whom I could talk about any of my concerns and he would always lead me to a solution.

I would like to express my thankfulness to every member of the Decision Systems & e-Service Intelligence Lab (DeSI) in the Centre for Artificial Intelligence (CAI). It was a wonderful experience to spend four years with these dedicated researchers. I will also have fond memories of UTS and Sydney. I especially thank Dr. Junyu

Xuan who helped me greatly during my PhD candidature, Dr. Vahid Behbood who strictly helped me solve several problems related to my research, Dr. Hua Zuo who was like an older sister to me and Dr. Qian Zhang who studied with me throughout the four years. I would also thank Dr. Fan Dong, Yiliao Song and Feng Liu who helped me and shared my joys and sadness.

I am incredibly grateful to my parents for their generosity and encouragement. This thesis would never have been completed without their constant support and understanding. I am also thankful to my friends who accompanied me and were by my side, listening to me and supporting me through the arduous journey of four years.

Abstract

Network representation aims to learn a latent feature space so that artificial intelligent algorithms can be applied based on the latent features. The set of latent features is obtained from the information hidden behind network structures, which is learned to provide knowledge for traditional machine learning tasks, such as node classification, recommendation and data visualization. Networks, which are a kind of structured data, limit the representation performance in the structure searching process. Therefore, a good node sampling strategy plays an important role in network representation. Recent research has driven significant progress in network representation by employing random walk as the network sampling strategy. However, real-world large-scale information networks naturally have structural sparsity. The existing approaches to random walk-based network representations are in the domain-specific view to represent the nodes in a vector format, which cannot guarantee a good representation by one network knowledge learning.

To address these gaps, this research proposes a framework and develops two algorithms to adapt useful information across relational large-scale information networks and allows the information of the network structure to be transferred from one network to another network to improve the performance of network representation. First, a novel framework of transferring structures across large-scale information

networks (FTLSIN) is proposed. FTLSIN consists of a two-layer random walk to measure the relations between two networks and predict the links across them. Second, a cross-domain network representation algorithm (CDNR) is proposed to demonstrate the knowledge which transfers across domains. CDNR learns the structural information from dense networks to sparse networks and further defines the two-layer random walk in unsupervised feature learning with a cross-domain node mapping procedure and a cross-domain walk mapping procedure. Thirdly, a cross-domain similarity learning algorithm (CDSL) is proposed to acquire the most relevant knowledge from the external network. CDSL is nested in the biased random walk-based node sampling and targets the minimum cost of searching the neighborhood in the biased random walk that considers the first-order and second-order walking; and the neighborhood is described by a dual centrality indicator which consists of closeness centrality and betweenness centrality. The developed framework and the two algorithms are very innovative and significantly contribute to both fields of transfer learning and network representation.

Table of Contents

CERTIFICATE OF AUTHORSHIP/ORIGINALITY	ii
Acknowledgements	iii
Abstract	v
List of Figures	xi
List of Tables	xiii
1 Introduction	1
1.1 Background	1
1.2 Research Questions and Objectives	5
1.2.1 Research Questions	5
1.2.2 Research Objectives	8
1.3 Research Contributions	10
1.4 Research Significance	12
1.5 Thesis Structure	13
1.6 Publications Related to this Thesis	17

2	Literature Review	20
2.1	Random Walk-based Network Representation	20
2.1.1	Network Representation for Large-scale Information Networks	21
2.1.2	Shortest Path-based Network Patten Mining	23
2.1.3	Network Representation by Random Walks	24
2.2	Graph-theoretic Node Importance Mining in Information Networks .	25
2.2.1	Graph-theoretic Indicators of Node Importance	29
2.2.2	Graph-theoretic Node Importance Mining on Network Topologies	31
2.2.3	Graph-theoretic Node Importance Mining on Transmission Mechanisms	38
2.2.4	Comprehensive Analysis and Findings	40
2.3	Transfer Learning	42
2.3.1	Transfer Learning Concepts	44
2.3.2	Transfer Learning using Naïve Bayes	48
2.3.3	Comprehensive Analysis and Findings	51
3	Framework of Transferring Structures across Large-scale Information Networks	53
3.1	Introduction	53
3.2	Problem Statement	54
3.3	Large-scale Information Network Structures Transfer Framework . .	55
3.3.1	Skip-gram in FTLSIN	55
3.3.2	Two-layer Random Walk in FTLSIN	57
3.4	Experiments	62

3.4.1	Datasets	62
3.4.2	Setups	63
3.4.3	Baselines	63
3.4.4	Parameters Setting	66
3.4.5	Result Analysis	67
3.5	Summary	73
4	Cross-domain Network Representations based on Random Walk Transfer	74
4.1	Introduction	74
4.2	Problem Statement	76
4.3	CDNR with CD2LRW	77
4.3.1	Random Walk Sampling Strategies: Domain-specific	79
4.3.2	Power-law Distribution: The Assumption	80
4.3.3	Bottom-layer Random Walk: Knowledge Preparation	81
4.3.4	Top-layer Random Walk: Knowledge Transfer	82
4.3.5	Top-layer Feature Learning: Knowledge Representation	86
4.4	Experiments	87
4.4.1	Datasets	88
4.4.2	Setups	96
4.4.3	Baselines	97
4.4.4	Parameters Setting	98
4.4.5	Result Analysis	99
4.5	Summary	107

5	Cross-domain Similarity Learning based on Network Patterns	108
5.1	Introduction	108
5.2	Problem Statement	110
5.2.1	Node Centralities	111
5.2.2	Cross-domain Network Representations	111
5.3	Cross-domain Similarity Learning based on Network Patterns	113
5.3.1	Dual Centrality	114
5.3.2	Dual Centrality based Biased Random Walk	115
5.3.3	Dual Centrality based Randomized Shortest Path	117
5.3.4	Algorithm of DCBRW-based CDSL	118
5.4	Experiments	119
5.4.1	Datasets	119
5.4.2	Baselines	123
5.4.3	Setups	124
5.4.4	DCBRW Parameter γ -Learning	125
5.4.5	Result Analysis	125
5.5	Summary	130
6	Conclusion and Future Research	131
6.1	Conclusions	131
6.2	Future Study	133
	Bibliography	136
	Abbreviations	159

List of Figures

1.1	An illustration of different networks with the same substructure . . .	2
1.2	Thesis structure	16
2.1	An illustration of information network with highlighted important node	26
2.2	Two network examples with different node importances	31
2.3	Graph-theoretic node importance mining methods on network topology	32
3.1	An illustration of FTLSIN	56
3.2	Power-law distributions of the networks	64
3.3	Power-law distributions of the random walks	65
3.4	Illustrations of 2-dimensional network representation by FTLSIN . .	69
3.5	Illustrations of 2-dimensional network representation by PCA . . .	70
3.6	Illustrations of 2-dimensional network representation by LLE	71
3.7	Illustrations of 2-dimensional network representation by Laplacian .	72
4.1	An illustration of CDNR in four steps	78
4.2	Power-law distributions of the Blog3 network and its random walks	90
4.3	Power-law distributions of the YouTube network and its random walks	91
4.4	Power-law distributions of the Facebook network and its random walks	92

4.5	Power-law distributions of the PPI network and its random walks . . .	93
4.6	Power-law distributions of the arXivCit-HepPh network and its random walks	94
4.7	Power-law distributions of the arXivCit-HepTh network and its random walks	95
5.1	An illustration of CDNR in FTLSIN	112
5.2	An example of DCBRW	115
5.3	LesM γ -learning results: $\gamma^* = 2.0423E - 01$	126
5.4	Facebook γ -learning results: $\gamma^* = 5.1496E - 04$	126
5.5	Blog3 γ -learning results: $\gamma^* = 1.8890E - 03$	127
5.6	PPI γ -learning results: $\gamma^* = 3.4623E - 03$	127
5.7	wiki γ -learning results: $\gamma^* = 1.4295E - 03$	128

List of Tables

3.1	FTLSIN dataset statistics	62
3.2	FTLSIN classification results on target domain network of M10 . . .	68
4.1	CDNR dataset statistics	88
4.2	CDNR domain selections based on network statistics	97
4.3	CDNR classification results of Micro-F1 on the target domain network of PPI	100
4.4	CDNR classification results of Macro-F1 on the target domain network of PPI	101
4.5	CDNR classification results of Micro-F1 on the target domain network of Facebook	102
4.6	CDNR classification results of Macro-F1 on the target domain network of Facebook	102
4.7	CDNR classification results of Micro-F1 on the target domain network of YouTube	103
4.8	CDNR classification results of Macro-F1 on the target domain network of YouTube	103
4.9	Pairwise t-test results of FTLSIN/CDNR versus baselines	106

5.1	CDSL dataset statistics	121
5.2	CDSL classification results on the target domain network of LesM .	129