

**EARLY IDENTIFICATION OF EXPERTS IN ONLINE  
QUESTION ANSWERING COMMUNITIES USING  
NEURAL NETWORKS**

J.H.M.M.D Herath

229330E

Degree of Master of Science

Department of Computer Science and Engineering

University of Moratuwa

Sri Lanka

March 2024

## **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).”

Signature:

Date: 03/06/2024

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

Signature of the supervisor:

Date 03/06/2024

## ABSTRACT

Community Question Answering (CQA) platforms designed to enable users to ask questions and seek answers from the community have gained an increased popularity over the years. Prior research has shown that these communities thrive largely due to a small subset of expert users who provide comprehensive and often accurate answers. Identification of such experts in community question answering (CQA) platforms is an important, yet challenging task.

Many prior studies employ variants of classical machine learning and link analysis techniques to tackle this challenging task. However, studies leveraging novel advancements in neural networks to tackle this challenging problem are quite sparse. This thesis work models community question answering forums as heterogeneous graphs and leverages novel Graph Neural Networks to learn rich representations of users. Next, this thesis work adapts a neural network model to leverage learned user representations and embeddings of questions extracted from a pre-trained large language model (LLM) to identify expert users who are best suited to answer a given question. Through experiments conducted on a real-world CQA dataset, the author demonstrates the effectiveness of the proposed approach.

Furthermore, existing studies have not sufficiently explored the effects of considering different degrees of relevance among neighbors in CQA interaction graph, and largely assume all neighbors to be of equal relevance. This thesis work explores this under-explored area, and experiments with Graph Attention Networks (GAT) which makes use of attention mechanisms to weigh the messages passed from neighboring nodes. Furthermore, this thesis work proposes a novel edge-weighting scheme based on temporal decay to explicitly assign relevance scores to interactions among nodes. Through experiments, the author demonstrates how capturing different degrees of relevance among neighbors helps improve the identification of community experts.

Keywords : Expert Identification, Expert Recommendation, Neural Networks, Graph Neural Networks, Community Question Answering

## **ACKNOWLEDGMENTS**

I would like to extend my utmost gratitude to my supervisor, Dr. Charith Chithraranjan for all the invaluable support given to make this research project a success.

## TABLE OF CONTENTS

DECLARATION.....	i
ABSTRACT.....	ii
ACKNOWLEDGMENTS.....	iii
TABLE OF CONTENTS.....	iv
LIST OF FIGURES.....	vi
LIST OF TABLES.....	vii
LIST OF ABBREVIATIONS.....	vii
LIST OF APPENDICES.....	vii
1. INTRODUCTION.....	1
1.1 Introduction to Community Question Answering Platforms.....	1
1.2 Expert Users in CQA Platforms and the Importance of Identifying Them.....	2
1.3 Challenging Nature of the Problem of Identification of Community Experts..	3
1.4. Research Problem.....	4
1.5. Research Objectives.....	5
2.LITERATURE REVIEW.....	6
2.1 Link Analysis Techniques.....	6
2.1.1 Leveraging Basic Graph Metrics.....	7
2.1.2 Variants of the PageRank Algorithm.....	7
2.1.3 Variants of the HITS algorithm.....	8
2.2 Content Based Techniques.....	9
2.2.1 Basic Approaches.....	9
2.2.2 Leveraging Probabilistic Language Models.....	10
2.2.3 Topic Models.....	12
2.2.4 Classification Techniques.....	13
2.3 Hybrid Approaches Extending Content-Based and Link-Analysis Techniques.	15
2.3.1 Approaches Adopting Link Analysis Techniques and Topic Models....	16
2.3.2 Approaches Adopting Link Analysis Techniques and Content Based Techniques.....	17
2.4 Deep-learning Based Techniques.....	17
2.5 CQA Datasets.....	19
2.5.1 StackOverflow Data.....	19
2.5.2 Yahoo! Answers Data.....	20
2.5.3 Other Data Sources.....	20
3. METHODOLOGY.....	21
3.1 Problem Statement.....	21

3.2 Dataset.....	22
3.2.1 Exploring the Dataset.....	23
3.2.2 Pre-Processing the Dataset and Feature Engineering.....	26
3.3 Modeling CQA Forums as Graphs.....	27
3.3.1.1 Homogeneous Graphs.....	28
3.3.1.2 Heterogeneous Graphs.....	28
3.4 Models.....	30
3.4.1.1 Training the GNN Model on an Edge-Regression Task.....	32
3.4.2 Graph Convolutional Network Models.....	33
3.4.3 Graph Attention Network ( GAT ) Models.....	33
3.4.5 Ranking Model.....	36
3.5 Evaluation Metrics.....	37
3.5.1 Precision@k.....	37
3.5.2 Mean Reciprocal Rank ( MRR ).....	37
3.6 Baseline Models.....	38
3.6.1 Latent Dirichlet Allocation (LDA) based Topic Model.....	38
3.6.2 Node2Vec.....	38
3.6.3 EndCold*.....	39
4. RESULTS AND DISCUSSION.....	40
4.1 Exploring the Research Question 1.....	40
4.1.1 Inspecting the Quality of the User Embeddings.....	44
4.2 Exploring the Research Question 2.....	46
4.3 Discussion.....	48
REFERENCES.....	50
APPENDIX A - StackExchange Data Explorer Query For Extracting the Initial Dataset.....	60
APPENDIX B - Snippets from the Code Related to Training Graph Neural Network Models.....	61
APPENDIX C - Snippets from the Code Related to the Training the Ranking Model..	63
APPENDIX D - Snippets from the Code Related to Visualizing the Learned Embeddings.....	65

## LIST OF FIGURES

	Page
Figure 3.1 An example question in StackOverflow along with associated question tags	24
Figure 3.2 An example accepted answer in StackOverflow and the score earned based on community votes	24
Figure 3.3 Distribution of the voting scores earned by answers in the training dataset	25
Figure 3.4 Distribution of the top tags in the dataset	26
Figure 3.5 A snapshot of the raw dataset with HTML fragments	26
Figure 3.6 User interaction graph based on question asking - answering behavior	28
Figure 3.7 An Undirected Heterogeneous Network consisting of Answer, Tag and User nodes	29
Figure 3.8 Overall architecture of our approach	30
Figure 3.9 Leveraging message passing mechanism in GNNs to learn suitable embeddings	31
Figure 3.10 The number of c# and kotlin answers given by the user 5133585 over time	35
Figure 4.1 Comparison of Precision@k Scores of the Models	43
Figure 4.2 Comparison of Mean Reciprocal Ranks ( MRR ) of the Models	43
Figure 4.3 Visualizing the Learnt User Embeddings	45
Figure 4.4 A 2D projection of a sample of user embeddings along with the most frequent question tags associated with each user	46

## LIST OF TABLES

Table 4.1 Results of the Experiments	41
--------------------------------------	----

## LIST OF ABBREVIATIONS

Abbreviation	Description
CQA ( Platforms )	Community Question Answering Platforms
GNN	Graph Neural Network
LLM	Large Language Model
GAT	Graph Attention Networks
GCN	Graph Convolutional Networks
NLP	Natural Language Processing
LDA	Latent Dirichlet Allocation
MRR	Mean Reciprocal Rank

## LIST OF APPENDICES

APPENDIX A - StackExchange Data Explorer Query For Extracting the Initial Dataset	60
APPENDIX B - Snippets from the Code Related to Training Graph Neural Network Models	61
APPENDIX C - Snippets from the Code Related to the Training the Ranking Model	63
APPENDIX D - Snippets from the Code Related to Visualizing the Learned Embeddings	65