



Information Systems Department

Faculty of Computer and Information Sciences

Ain Shams University

Efficient Hybrid Technique for Community Question Answering

Thesis submitted as a partial fulfillment of the requirements for the degree of
Master of Science in Computer and Information Sciences

By

Dalia Magdy Mohamed Talaat El Alfy

Teaching Assistant at Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University

Under Supervision of

Prof. Dr. Khaled Bahnasy

Professor at Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University

Dr. Rasha Mohamed Ismail

Associate Professor at Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University

Dr. Walaa Khaled Gad

Associate Professor at Information Systems Department,
Faculty of Computer and Information Sciences, Ain Shams University

Acknowledgment

First and above all, I want to thank Allah the almighty for providing me the opportunity to proceed successfully and the capability to overcoming any challenges and obstacles that have faced me during the preparation of this thesis. This thesis appears in its current form due to the assistance and full support from many people. Therefore, I would like to offer my gratitude and sincere thanks to all of them.

I would like to acknowledge and express my deep thanks to Prof. Dr. Khaled Bahnasy for the continuous encouragement, valuable guidance and his leadership throughout my years of study and through the process of researching and writing this thesis.

I would like to express my deepest gratitude and appreciation to Dr. Rasha Ismail for her advices, caring, encouragement and her great valued and supportive supervision.

Also, I would like to sincerely and deeply thank Dr. Walaa Gad for her patience, understanding, enthusiasm, inspiring instructions and great help that motivates me during rough moments throughout this thesis.

A special appreciation, many thanks and a great love goes to my family for being always for me and their unfailing support. First of all, I want to warmly thank my father's pure soul. Thank you for every moment we passed together and every word you teach me during our few years together. I want to grant him any success in my life as always being my inspiration idol. I also want to thank my mother for her prayers for me that made me sustained in these years. I cannot forget my sisters' spiritual support and prayers for me. My beautiful little niece, your smile wipes any feeling of tired, thank you for being you and for being in my life, god bless you. I would like to forward many thanks and appreciation to my grandmother's soul that encouraged me a lot during her life time.

Lastly, I won't forget to thank my husband for his unending support, continued encouragement and understanding during my research. Many thanks to you, I appreciate all these things. Without his spiritual support, I would not have been able to finish this research.

List of Publications

- D. Elalfy, W. Gad and R. Ismail, "Predicting best answer in community questions based on content and sentiment analysis", in IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2015, pp. 585-590.
- D. Elalfy, W. Gad and R. Ismail, "A hybrid model to predict best answers in question answering communities", Egyptian Informatics Journal, 2017, IF=1.77.

Abstract

Question answering communities (QAC) are nowadays becoming widely used due to the huge facilities and flow of information that it provides. These communities target is to share and exchange the knowledge between users. Through asking and answering questions under large number of categories.

Unfortunately, there are a lot of issues existing that made knowledge process became a difficult one. One of those issues is that not every asker has the knowledge and ability to select the best answer for his question, or even selecting the best answer based on subjective matters. The analysis in this thesis is conducted on stack overflow community. In this work, a hybrid model for predicting the best answer is proposed. The proposed model is consisting of two modules. The first module is the content feature which consists of three types of features: question-answer features, answer content features, and answer-answer features. In the second module, a novel reputation score function to stack overflow community is used as a non-content feature to predict the best answer. Then, a hybrid model is proposed that merge content and non-content models and use them in the prediction. Study conducted experiments to train three different classifiers using the new added features. The prediction accuracy in the content and the proposed hybrid model is 88.36 % and 88.65% respectively.

Table of Contents

Abstract	IV
List of Figures	VIII
List of Tables.....	IX
List of Abbreviations	X
List of Abbreviations	XI
Chapter 1: Introduction.....	2
1.1 Overview	2
1.2 Motivation.....	5
1.3 Objectives.....	5
1.4 Contributions.....	6
1.5 Thesis Organization	8
Chapter 2: Literature Survey	11
2.1 Introduction.....	11
2.2 Recommend Experts to the Current Question.....	11
2.3 Finding the Best Answer.....	15
2.4 Finding Collaborative Experts.....	22
2.5 Route Questions to a Specific Expert	22
Chapter 3: The Proposed Hybrid Model Architecture	24

3.1	Introduction.....	24
3.2	Content based module Architecture.....	24
3.2.1	Preprocessing Module.....	25
3.2.2	Features Selection Model.....	28
3.2.3	Classification Process	30
3.3	Non-Content based Model Architecture.....	40
3.3.1	Reputation Based Non-Content Model.....	40
3.3.2	Participation level.....	41
3.3.3	Best Answer Level	41
3.3.4	Expertise Level.....	42
3.3.5	Confidence level	42
3.3.6	Reputation Score	43
3.4	The Proposed Hybrid Model	43
Chapter 4: Experiment and Results.....		45
4.1	Dataset Description	45
4.2	System Description.....	45
4.3	Performance Measures	46
4.4	Content Model Experiment.....	47
4.5	Content Features Results	47
4.6	Local Reputation score Model Results (RP).....	52
4.7	Hybrid Model Experiments.....	54

4.7.1 HCR Model	54
4.7.2 SR Model	56
4.7.3 HSR Model	58
4.7.4 Separated Non-Content Model (SP)	60
4.7.5 HSP Model	62
4.8 Comparing Results and Evaluation	65
Chapter 5: Conclusion and Future Work	69
5.1 Conclusion	69
5.2 Future Work	70
Appendix A: Calculate the Standard Deviation	71
Appendix B: Convert HTML into Text	74
References	90

List of Figures

<i>Figure 3.1 Content Based Model Architecture</i>	25
<i>Figure 3.2 Example of how Random Forest classifier works</i>	36
<i>Figure 3.3 Random forest tree</i>	37
<i>Figure 3.4 All trees of random forest</i>	38
<i>Figure 3.5 User's Reputation Score</i>	40
<i>Figure4.1 Content Classifiers Results</i>	51
<i>Figure 4.2 Local Reputation Classifiers Results</i>	54
<i>Figure 4.3 HCR Model Results</i>	56
<i>Figure 4.4 SR Model Results</i>	58
<i>Figure 4.5 HSR Model Accuracy</i>	60
<i>Figure 4.6 SP Model Results.</i>	62
<i>Figure4.7 HSP Model Results</i>	64

List of Tables

<i>Table 2.1 Recommend Experts to a Specific Question.....</i>	<i>15</i>
<i>Table 2.2 Finding the Best Answers.....</i>	<i>21</i>
<i>Table 3.1 A sample of HTML dataset used.....</i>	<i>26</i>
<i>Table 3.2 A sample of the used Stop Words in the dataset.....</i>	<i>27</i>
<i>Table 4.1 Features designed under each type and their description.....</i>	<i>48</i>
<i>Table 4.2 Content based model results.....</i>	<i>49</i>
<i>Table 4.3 Best answer prediction accuracy in RP Model.....</i>	<i>53</i>
<i>Table 4.4 Best answer prediction accuracy after adding content and non-content features in classifier (HCR).....</i>	<i>55</i>
<i>Table 4.5 Best answer prediction accuracy using non-content feature in classifier SR.....</i>	<i>57</i>
<i>Table 4.6 Best answer prediction accuracy after adding content and stack overflow non-content feature in classifier.....</i>	<i>59</i>
<i>Table 4.7 Best answer prediction accuracy using non-content separated features in classifier.....</i>	<i>61</i>
<i>Table 4.8 Best answer prediction accuracy using both content and non-content separated features in classifier.....</i>	<i>63</i>
<i>Table 4.9 Comparison between the three non-content models in prediction accuracy.....</i>	<i>65</i>
<i>Table 4.10 Comparing three hybrid models in prediction accuracy.....</i>	<i>66</i>

List of Abbreviations

CQA: Collaborative Question Answer Community.

ESN: Enterprise Social Network.

ASG: Aggregate Specialization Graph.

SSG: Specialization Sub Graph.

LDA: Latent Dirichlet Allocation Model.

TF: Term Frequency.

IDF: Inverse Document Frequency.

QA: Question Answer community.

MOOC: Massive Open Online Course.

SF: Server Fault Community

CO: Cooking Community.

CHPT: Convert HTML into Plain Text.

TP: Tokenization Process.

RSW: Remove Stop Words.

AWD: Add Words into Dictionary.

F_A: Content Feature.

F_{A-A}: Context Feature.

F_{Q-A} : Question-Answer Feature.

List of Abbreviations

T_{PS}: True **P**ositives.

F_{PS}: False **P**ositives.

F_{NS}: False **N**egatives.

T_{NS}: True **N**egatives.

RF: Random **F**orest.

LR: Logistic **R**egression.

NB: Naïve **B**ayes.

RNC: Reputation **N**on-Content model.

HCR: Hybrid **C**ontent and **R**eputation model.

SR: Stack overflow **R**eputation model.

HSR: Hybrid content and **S**tack's **R**eputation model.

SP: Se**P**arated non-content model.