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

List of Figures

List of Tables

List of Abbreviations

CQA: **C**ollaborative **Q**uestion **A**nswer Community.

ESN: **E**nterprise **S**ocial **N**etwork.

ASG: **A**ggregate **S**pecialization **G**raph.

SSG: **S**pecialization **S**ub **G**raph.

LDA: **L**atent **D**irichlet **A**llocation Model.

TF: **T**erm **F**requency.

IDF: **I**nverse **D**ocument **F**requency.

QA: **Q**uestion **A**nswer community.

MOOC: **M**assive **O**pen **O**nline **C**ourse.

SF: **S**erver **F**ault Community

CO: **Co**oking Community.

CHPT: **C**onvert **H**TML into **P**lain **T**ext.

TP: **T**okenization **P**rocess.

RSW: **R**emove **S**top **W**ords.

AWD: **A**dd **W**ords into **D**ictionary.

F $_{A}$: Content **F**eature.

F $_{A\text{-}A}$: Context **F**eature.

F $_{Q\text{-}A}$: **Q**uestion-**A**nswer **F**eature.

List of Abbreviations

T_{PS}: **T**rue **P**ositives.

F_{PS}: **F**alse **P**ositives.

F_{NS}: **F**alse **N**egatives.

T_{NS}: **T**rue **N**egatives.

RF: **R**andom **F**orest.

LR: **L**ogistic **R**egression.

NB: **N**aïve **B**ayes.

RNC: **R**eputation **N**on-**C**ontent model.

HCR: **H**ybrid **C**ontent and **R**eputation model.

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

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

SP: **S**e**P**arated non-content model.