NTNU
Norwegian University of
Science and Technology

Question analysis of coding questions on StackOverflow

130533 - Knut Lucas Andersen

31-05-2016

Master's Thesis
Master of Science in Applied Computer Science
30 ECTS
Department of Computer Science and Media Technology
Norwegian University of Science and Technology,

Supervisor 1: Assoc. Prof. Simon McCallum
Supervisor 2:

Preface

This is my Master thesis concluding the two years spent at NTNU Gjøvik: Master Applied Computer Science - Web, Mobile, Games track. The thesis was carried out during the spring semester 2016, from January to the end of May.

The main concept for the thesis was based on discussions with supervisor. The original plan was to create a Chat Agent that could answers students questions and give feedback to their question quality, by using StackOverflow as a knowledge base. However, during the Master thesis project presentation, other professors noted that the scope of the project was to large for a Master thesis. The thesis were therefore narrowed down to focus on coding questions posted on StackOverflow, in an attempt to evaluate question quality and predict the future votes for a given question.

31-05-2016

Acknowledgment

I would like to thank the following persons for their help and support during all these years. It would not have been possible without them.

My supervisor, Simon McCallum, for understanding my difficult situation and for his patience and helpful advices on how to proceed so that I could complete my Master thesis.

Mariusz Nowostawski for his advice on how to get started with text processing and ideas for building the SVM model.

My best friend, Njål Dolonen, for always being there for me, and ensuring that I work by checking up on me.

My grandmother, Mimmi H. Underland, may she rest in peace. None of this would have been possible without your support, understanding, love and care. This is for you.

I would also thank my family, for believing in me and supporting me through this.

K.L.A

Abstract

This is a summary of the thesis including the conclusion of what was discovered.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Topic covered/Research area

The goal of this thesis is to research and analyse coding questions posted on StackOverflow. Since most systems developed and researched focuses on finding good answers, it would be interesting to see if it was possible to develop a system that could analyse and predict good questions. To narrow down the field, the focus here will only be on programming and coding questions. Therefore StackOverflow was selected. StackOverflow is a part of the StackExchange community, where each branch is related to a specific field with domain experts. Questions (and answers) can be ranked high or low by being given votes. Questions can also be closed for various reasons (e.g. duplicate, off topic, unclear, etc) [1].
To be able to analyse questions, an Artificial Intelligence (AI) system was developed by using the Support Vector Machines (SVM) library LIBSVM. The data used for the training was XML dumps of the StackOverflow database [2].
. . . write more here. . .

1.2 Problem description

Question-answering is today a field where the main focus is on finding answers that best matches the given question, regardless of the quality of the question asked. However, there is not a lot of research being done related to how to improve the quality of the question people asks. In both higher education (Bachelor, Master and Ph.d.) and research, the first step is to define one or more questions to describe the problem being solved. For students (and researchers), this can easily become a tedious task. What if there was a system that could give a prediction on the quality of your question?

1.3 Research questions

- What defines a good (coding) question?
- Can we predict if a new question will be defined as a good question by use of SVM?
- If a question has a lot of good answers, does this influence the rating of the question? (e.g. question-answers with well formatted code snippets)
- Can the results of the SVM be improved if the feature set is using more complex values[1]?

[1] Perhaps not the best formulation, but basically the question is whether or not there would be any improvement to the analysis if one added additional measurements. E.g. if de-facto is to

1.4 Methodology to be used

1.5 Justification, Motivation and Benefits

One can draw comparison between academic peer-review and the peer-review done within the StackExchange community. The posted content is read, scored and given feedback. Based on that feedback, users can then choose to update or edit their question (or answer). Therefore, using both the data and the answers as resources are completely valid. LIBSVM is a library that has been used both in research and education, and there should therefore be no problems with presented results and its quality (*will add references here to papers using libsvm*). Using sigmoid threshold because the interest lie within whether or not it is a good question, and not the classification of why it is a bad question.
There is not a lot of research or resources related to asking and defining good questions. With this research, a system can be developed that can not only give your question a score, but in the future also point out which parts of the question are bad. The system can be used both by researchers and students to improve their question quality and learn to be better at asking questions. This in turn will be beneficial for the work/research they are doing.

1.6 Limitations

Amount of time available; e.g. processing of questions from dataset. libsvm is large and thus would require a lot of time if fine tuning were needed. not all the available data from the dataset can be tested

1.7 Thesis contribution

Research what the users of StackOverflow sees as good questions and use this to create a system for predicting good coding questions.

1.8 Thesis structure

summary of the thesis; e.g. in Chapter 2 previous research is presented, in Chapter 6 a conclusion is made.

only use question sentence, what happens if length and symbol (e.g. smileys, question marks, etc.) measurements were added? Could it then predict more accurately whether or not this was a good question?

2 State of the art

basically a summary of existing and relevant research for this thesis

2.1 Text classification

a more general introduction to text classification and the AI techniques used
e.g. papers comparing systems to use as argument for why svm was chosen

2.2 Question-Answering (Q/A)

i'm thinking this needs to be split into sub-subsections or several sub-sections, mainly because I'm a bit unsure what I should put here
the following are some examples:

- question-answering techniques and analyses
- ontology and semantics
- StackExchange and other online communities for qa (potentially separate section)

2.3 Support Vector Machines (SVM) and LIBSVM

basically papers that have used (lib)svm for text classification, qa, etc

3 Methodology

3.1 Support Vector Machines (SVM)

this might either be a) removed or b) merged with section below

3.2 LIBSVM Implementation

basically an explanation as to how libsvm was used in this thesis

3.3 Feature sets, attributes and processing

what features were selected and why?
how was data processed (e.g. retrieved from db, converted to libsvm format, and so forth)
attributes: length, symbols, question sentence only, code snippet, votes, closed, etc.

4 chapter4

empty placeholder section

5 Discussions

5.1 Data and Testing

discussion on the data set and how it was tested.
the results and what they showed. potential improvements, etc.

5.2 Artificial Intelligence (AI) Methods

alternative methods and options (e.g. one could have used ann or k-nn, but as shown in. . .)

5.3 Implementation Architecture

discussion on the code that was written and its functionality
what worked, what should be updated/changed, etc.

6 Conclusion/Summary

6.1 Overview of main results

basically what the title says; a summary of the results

6.2 Further work

additions and updates to the system
new research possibilities based on results

Bibliography

[1] Stackoverflow.com. What does it mean if a question is "closed" or "on hold"? (online). 2016. URL: http://stackoverflow.com/help/closed-questions.

[2] StackExchange, I. Stack exchange data dump (online). 2016. URL: https://archive.org/details/stackexchange.

A Appendix

A.1 Acronyms

AI Artificial Intelligence. 1, 3

SVM Support Vector Machines. 1

A.2 Data sets/Statistical Overview