UNIVERSITI PUTRA MALAYSIA

EFFECTS OF EXTENDED FEATURES ON TEXT-BASED CLASSIFIER FOR CORRECTIVE MAINTENANCE

NAGHMEH MAHMOODIAN

FSKTM 2011 22

EFFECTS OF EXTENDED FEATURES ON TEXT-BASED CLASSIFIER FOR CORRECTIVE MAINTENANCE

By

NAGHMEH MAHMOODIAN

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfillment of the Requirement for the Degree of Master of Science

June 2011

DEDICATION

To my parents and my husband and my daughter

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Master of Science

EFFECTS OF EXTENDED FEATURES ON TEXT-BASED CLASSIFIER FOR CORRECTIVE MAINTENANCE

By

NAGHMEH MAHMOODIAN

June 2011

Chairman: Rusli Abdullah, PhD

Faculty: Computer Science and Information Technology

Software maintenance (SM) is a complex process and is composed of various tasks that are supported by software maintainer. Classification of maintenance request (MR) is one of the tasks in large software system, yet it is often not well classified. Classification of the MRs depends on their types, which are corrective, adaptive, perfective or preventive, which are also known as maintenance type (MT). The MTs are important in keeping the quality factors of the software system. Especially, corrective maintenances are the most requests which are released in bug tracking system (BTS) in comparison to other MTs. Corrective maintenance indicates the modification of software product after its delivery in

order to correct the discovered faults, and non-corrective indicated other types of maintenance.

However, classification of MT is difficult in nature and this affect maintainability of the system. A number of researches in this area are dedicated to automate methods and processes in SM in order to aid MT classification. Thus, there is a need for tools that support the maintainers in doing their daily maintenance activities more effectively. The tools should be able to classify MRs automatically without human intervention in performing MT classification.

MR is textual information that can be categorized according to various features by using machine learning (ML) techniques. Title, description, error encountered, and source of request are four features for the datasets which are used to train the classifiers. The two recent features, *error encountered* and *source of request* are considerate as new features in this study. These new features are added to two previous features (title and description) which were used by Antoniol *et al*., (2008). The goal of this research is to increase the classification accuracy of MRs into MT by using these features and present the effect of each feature in determining MT and show the best feature among the two new features. Next, the textual information of the reported bugs in the BTS will be also considered to determine whether it is sufficient to classify the MRs into corrective or non-corrective MTs.

This research also presents the results of applied combining new features in the MR classification, which affect the maintainability and other quality factors of the software system. Three different classification techniques, namely Decision Tree, Naïve Bayesian, and Logistic Regression are used as the classifier.

The dataset used in the experiment are from three BTS, which are Mozilla, Eclipse and JBOSS. The dataset comprises of 1800 MRs with the corresponding features. The experimental results show that the proposed MRMT model is able to achieve higher classification accuracy. The MRMT model, which is employed two more features, namely source of request and error encountered, has also outperformed the previous works.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

KESAN DARIPAPA CIRI-CIRI LANJUTAN TERHADAP PENGELAS BERASASKAN-TEKS UNTUK PENYELENGGARAAN PEMBETULAN

Oleh

NAGHMEH MAHMOODIAN

June 2011

Pengerusi: Rusli Abdullah, PhD

Fakulti: Sains Komputer dan Teknologi Maklumat

Penyelenggaraan perisian (SM) merupakan satu proses yang kompleks dan terdiri daripada pelbagai tugasan yang disokong oleh penyelenggara perisian. Pengklasifikasian permintaan penyelenggaraan (MR) adalah salah satu daripada satu tugasan dalam sistem perisian yang besar, namun ianya masih belum dapat diklasifikasikan dengan baik. Pengklasifikasian MR bergantung kepada jenis penyelenggaraan, iaitu sama ada penentuan ralat, adaptasi, kesempurnaan atau perlindungan. Jenis-jenis penyelenggaraan, dan dipanggil sebagai jenis-jenis penyelenggaraan (MT). MT adalah penting dalam memelihara faktor kualiti dalam sesebuah sistem perisian. Terutamanya, penyelenggaraan pembetulan merupakan permintaan yang paling tinggi di mana ia diperkenalkan dalam sistem penjejakan pepijat jika dibandingkan dengan jenis-jenis penyelenggaraan (MT)

yang lain. Penyelenggaraan pembetulah menunjukkan piubahan pada produk pirisian selepas penyampaiannya bagi membetulkan kesalahan yang ditemui, manakala bukan-pembetulan menunjukkan jenis penyelenggaraan yang lain.

Walau bagaimanapun, pengklasifikasian MT adalah sulit secara semulajadi dan ini memberi kesan ke atas penyelenggaraan sistem. Terdapat cabang penyelidikan yang besar bagi mengautomasikan kaedah dan proses SM bagi membantu pengklasifikasian MT. Oleh yang demikian, timbul keperluan alatan-alatan yang menyokong para penyelenggara menjalankan aktiviti penyelenggaraan harian mereka. Alatan-alatan tersebut perlu berupaya mengkategorikan MR secara automatik tanpa interaksi dengan manusia sewaktu menjalankan pengklasifikasian MT.

Punca-punca MR ada maklumat berasaskan teks yang boleh dikategorikan mengikut pelbagai ciri dengan menggunakan teknik-teknik pembelajaran mesin (ML). Tajuk, penerangan, penemuan ralat, dan sumber permintaan merupakan empat ciri bagi set data di mana ciri-ciri tersebut digunakan untuk melatih pingilas, pinimuah ralat dan sumbir pirmintaan dianggap sibagai dua civi tirbaru dalam kajian ini. Dua ciri baru ini ditambah pada dua ciri yang terdahulu (tajuk dan penerangan) yang mana telah digunakan oleh Antoniol et al., (2008). Matlamat kajian ini adalah untuk meningkatkan ketepatan pengklasifikasian MR ke dalam MT. dengan menggunakan ciri-ciri tersebut dan mempersembahkan

kesan daripada ciri-ciri tersebut dalam menentukan MT dan sekaligus minunjukkan ciri yang terbaik diantara kedua-dua ciri baru tersebut. Kemudian, teks laporan ralat daripada sistem mengesan ralat (BTS) digunakan bagi menentukan sama ada ianya mencukupi untuk mengklasifikasikan MR ke dalam MT pembetulan atau bukan pembetulan.

Kajian ini juga mempersembahkan keputusan pengaplikasian ciri-ciri baru sewaktu pengklasifikasian MR ke dalam MT, yang memberi kesan ke atas penentuan penyelenggaraan dan faktor kualiti dalam sesebuah sistem perisian. Tiga teknik pengklasifikasian, iaitu Decision Tree, Naive Bayesian dan Logistic Regression telah diaplikasikan untuk menjalankan pengklasifikasian.

Data yang digunakan dalam eksperimen adalah diambil daripada tiga BTS, iaitu Mozilla, Eclipse dan JBOSS. Set data yang digunakan adalah terdiri daripada 1800 MR dengan ciri masing-masing. Keputusan eksperimen menunjukkan bahawa model MRMT yang dicadangkan berupaya untuk mencapai ketepatan yang lebih tinggi dalam pengklasifikasian MR ke dalam MT. Model MRMT tersebut juga telah menandingi keputusan daripada kajian lepas melalui atribut sumber permintaan dan ralat yang dijumpai.

ACKNOWLEDGEMENTS

First of all, I am thankful to God for giving me the opportunity to discover myself through life in Malaysia.

I would like to thank my supervisor, Dr.Rusli Abdullah for his valuable comments and advice through the course of this research. His encouragement and professional insights helped to shape this thesis and the research behind it. I would also like to thank my co-supervisor, Dr. Masrah Azrifah Azmi Murad for her advices and comments, which guides me through proper direction.

Also, my eternal gratitude is owed to my family who have been supportive in everything I have done. In particular, I would like to thank my mother and father, Shahin and Ahmad for their never ending love and support. I am highly indebted to my husband, Sasan for his understanding, encouragement and support throughout my study.

APPROVAL

I certify that an Examination Committee has met on **date of viva** to conduct the final examination of **Naghmeh Mahmoodian** on her **Master of Science** thesis entitled **"EFFECT OF EXTENDED FEATURES ON TEXT-BASED CLASSIFIER FOR CORRECTIVE MAINTENANCE"** in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1971 and Universiti Putra Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Abdul Azim b Abd Ghani, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Azreen bin Azman, PhD
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Muhamad Taufik bin Abdullah, PhD
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Wan Mohd. Nasir Wan Kadir, PhD
Associate Professor
Software Engineering Department
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
(External Examiner)

NORITAH OMAR, PhD
Associate Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirements for the degree of Master of Science. Members of the Supervisory Committee were as follows:

Rusli Abdullah, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Masrah Azrifah Azmi Murad, PhD
Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

HASANAH MOHD GHAZALI,PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

DECLARATION

I declare that the thesis is my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously, and is not concurrently, submitted for any other degree at Uuniversiti Putra Malaysia or at any other institutions.

_____

NAGHMEH MAHMOODIAN
Date: 17 June 2011

TABLE OF CONTENTS