Relation Extraction in Term Weighting for Text Classification

Dat Tan Huynh

(in Vietnamese: Huỳnh Tấn Đạt)

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Master of Information Sciences and Engineering
(by research)

University of Canberra

December, 2010

University of Canberra

Abstract

Relation Extraction in Term Weighting for Text Classification

Dat Tan Huynh
(in Vietnamese: Huỳnh Tấn Đạt)

With the explosion of WWW information and the increasing availability of documents in digital form, information acquisition and organisation needs are becoming more and more significant. The automated text classification task, that is how to categorise documents into pre-defined categories, has been a matter of major concern. A typical text classification (TC) system has two main components, namely document representation and classification. Some approaches to document representation have been used term frequency (TF), inverse document frequency (IDF), term category dependency (TCD) and term co-occurrence (TCO). Recent approaches to document representation look for semantic relationships among terms. The relations can be extracted from particular collections such as Wikipedia.

Although these popular approaches to document representation have been successful in text classification, TF and IDF are struggling to differentiate documents and cannot achieve high classification results with some abstract and complex corpora. TCO-based methods use term relations, however relations in terms of semantic aspects are still open questions. In summary, the problems raised below are the particular issues examined in this thesis:

1. How to extract potential relationships among terms to provide additional information for document representation?

2. How to take the advantages of the extracted relationships to build document representations?

3. How to consider category information to justify weightings of features, which can be used by popular text classifiers to leverage the classification results?

The thesis contributes the following solutions to solve the problems:

1. A Relation Extraction Framework: The goal of this framework is not only to present a method for extracting relationships among terms from a given text document, but also to present the preliminary work toward building a semantic relation extraction for TC. Moreover, it also shows an alternative choice of generating features from a text document instead of using the popular term-based approaches.

2. A Term Weighting Approach based on Relation Extraction and Graph Model: The aims of this method are not only to address the issues of popular term weighting approaches based on term frequencies and term co-occurrences, but also to present a way of talking the advantage of extracted relations in weighting terms. This work can be considered an example, and an opening to further investigation into applying other kinds of semantic relations, graph centrality ranking for various types of text classification tasks.

3. An Adaptable Term Weighting Framework for Text Classification: The framework shows how to apply the TCD measure in weighting terms for TC. The round-robin process from the framework helps to find out the suitable term weighting schema for the document based on category information.

Certificate of Authorship of Thesis

Except where clearly acknowledged in footnotes, quotations and the bibliography, I certify that I am the sole author of the thesis submitted today entitled

Relation Extraction Effects in Term Weighting for Text Classification

I further certify that to the best of my knowledge the thesis contains no material previously published or written by another person except where due reference is made in the text of the thesis.

The material in the thesis has not been the basis of an award of any other degree or diploma except where due reference is made in the text of the thesis.

The thesis complies with University requirements for a thesis as set out in Gold Book Part 7: Examination of Higher Degree by Research Theses Policy, Schedule Two (S2).

Refer to http://www.canberra.edu.au/research-students/goldbook

Signature of Candidate

Supervisory Panel:

_____

Signature of Chair of the Supervisory Panel: A/Prof. Dat Tran

Date: _____

ACKNOWLEDGMENTS

I am grateful to people who assisted me throughout the process of my research journey.

First, I am thankful for the generous support provided by my supervisor, Associate Professor Dat Tran. I am honoured to have a opportunity to work with him and thankful for his endless guidance, support, encouragement.

I would also like to express my sincere gratitude to supervisor Assistant Professor Wanli Ma, for his supports and feedbacks during the course of my study.

I am grateful to Professor Dharmendra Sharma and Professor John Campbell for their consistent help and support within the Faculty of Information Science and Engineering (ISE).

I am grateful Dr Nguyen Thai Son for the invaluable support and encouragement. I am honoured to have an opportunity to improve my research methodology in Australia. I specially acknowledge the University of Pedagogy in Ho Chi Minh City and Second Higher Education project in Vietnam for providing me financial support during my study.

I would like to thank everyone in Faculty of ISE–especially Hanh Huynh, Len Bui, Trung Le, Tuan Hoang, Phuoc Nguyen–for providing helpful assistances, discussions and productive collaborations.

Finally, I specially thank to Beth Barber for helping me enhance this thesis.

DEDICATION

to my parents

PUBLICATIONS

Huynh, D., Tran, D., Ma, W. and Sharma, D. (2011a), A New Term Ranking Method Based on Relation Extraction and Graph Model for Text Classification, *in Proc. of the* 34^{th} *Australasian Computer Science Conference (ACSC2011)*, Australian Computer Society.

Huynh, D., Tran, D., Ma, W. and Sharma, D. (2011b), Adaptable Term Weighting Framework for Text Classification, in *Proc. of the* 12^{th} *International Conference on Intelligent Text Processing and Computational Linguistics(CICLing2011)*, Springer Lecture Notes in Computer Science.

Huynh, D., Tran, D., Ma, W. and Sharma, D. (2011c), Grammatical Dependency-Based Relations for Term Weighting in Text Classification, in *Proc. of The* 15^{th} *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011)*.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

CRF: Category Relevant Factor

CVG: A method aims to detect convergence status

DF: Document Frequency

DIFF: A method to identify a number of differences between two label sets.

DIPRE: Dual Iterative Pattern Relation Extraction

DLA: Deep Linguistic Analysis

ESA: Explicit Semantic Analysis

FN: False Negative

FP: False Positive

GR: Gain Ratio

IDF: Inverse Document Frequency

IG: Information Gain

KNN: K-Nearest Neighbours

NSF: US National Science Foundation

OIE: Open Information Extraction

PR: A term ranking method based on relation extraction and Graph model

PR×IDF: A term weighting approach based on PR and IDF

PR×TCD: A term weighting approach based on PR and TCD framework

RF: Relation Frequency

SA: Surface Analysis

SVM: Support Vector Machine

TC: Text Classification

TCD: Term Category Dependency

TCO: Term Co-occurrence

TF: Term Frequency

TF×IDF: A term weighting approach based on TF and IDF

TF×TCD: A term weighting approach base TF and TCD framework

TP: True Positive

TN: True Negative