

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**

AUSTRALIA'S CAPITAL UNIVERSITY

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 34th 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 12th 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 15th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011)*.

TABLE OF CONTENTS

	Page
List of Figures	v
List of Tables	vii
Glossary	ix
Chapter 1: Introduction	1
1.1 Current Approaches for Document Representation	1
1.2 Problem Statements	3
1.3 Contribution of the Thesis	5
1.3.1 A Relation Extraction Framework	5
1.3.2 A Term Weighting Approach based on Relation Extraction and Graph Model	5
1.3.3 An Adaptable Term Weighting Framework for Text Classification	5
1.4 Extension of the Thesis	6
Chapter 2: Literature Review	7
2.1 Term Weighting Approaches	7
2.1.1 TF-based Approaches	9
TF Weighting Method	9
Normalised TF Weighting Method	9
Logarithm TF Weighting Method	9
2.1.2 IDF-based Approaches	10
IDF Weighting Method	10
TF×IDF Weighting Method	10
Normalised TF×IDF Weighting Method	11
2.1.3 Category-based Approaches	11

2.1.4	TCO-based Approaches	13
2.1.5	Relation-based Approaches	14
2.2	Relation Extraction Approaches	17
2.2.1	Surface Analysis Approaches	17
2.2.2	Deep Linguistic Analysis Approaches	19
Chapter 3:	Relation Extraction Method	21
3.1	Introduction	21
3.2	Relation Extraction Framework	21
3.2.1	Definition of Relation	21
3.2.2	A General Framework to Extract Relations	22
3.3	Linguistic Linkage Structures	24
3.4	Heuristic Algorithm	25
3.5	Relation Normalisation	27
3.6	Performances and Evaluations	27
3.6.1	Datasets	27
Wikipedia Corpus	28	
Ohsumed corpus	28	
NSFAwards corpus	28	
3.6.2	Relation Extraction Results	29
3.7	Summary	29
Chapter 4:	Term Weighting Approaches Based on Relation Extraction and Graph Model	31
4.1	Introduction	31
4.2	Graph Construction	32
4.3	Graph Ranking	32
4.4	Term Weighting Approach based on Graph Model	34
4.4.1	Term Weighting Measure Based on Graph Ranking and Inverse Document Frequency	34
4.4.2	Term Weighting Measure Based on Graph Ranking and Term Category Dependency	34
4.5	Summary	36

Chapter 5:	Adaptable Term Weighting Framework Based on Category Information	37
5.1	Introduction	37
5.2	Term Weighting Framework Based on Category Information	38
5.2.1	Training phase	40
5.2.2	Testing Phase	40
	Initialising Document Representation Process	40
	Round-robin Justifying Document Representation Process	41
5.2.3	How to Apply the Framework for Text Classification	41
5.3	Apply Term Weighting Framework for Text Classification Task	42
5.3.1	Term Weighting Method Based on Category Information	42
5.3.2	Convergence Status Checking	43
5.3.3	Label Set Aggregation	44
5.4	Experiments	45
5.4.1	Corpora	45
	Reuters-21578	45
	Ohsumed Corpus	46
	NSFAwards Corpus	46
5.4.2	Specified Term Weighting Framework for Text Classification Task	47
5.5	Performance & Discussion	48
5.6	Summary	50
Chapter 6:	Experiments and Discussions	53
6.1	Experiment Setups	53
6.1.1	Datasets	53
6.1.2	Classifiers	54
6.2	Comparative Methods	54
6.2.1	Term Weighting Method Based on TF and IDF	54
6.2.2	Term Weighting Methods Based on TCO Relations	56
6.2.3	Hybrid Methods Based on Term Category Dependency	57
6.3	Measurements	59
6.3.1	Precision and Recall	59
6.3.2	F_1 Measures	59

6.3.3	Micro-averaged and Macro-averaged Measures	60
6.4	Performances and Evaluations	61
6.4.1	Methodology and Features Summaries	61
6.4.2	Results from TCD-based Methods	62
6.4.3	Results from IDF-based Methods	64
6.4.4	IDF vs. TCD	65
6.5	Summary	66
Chapter 7:	Discussions and Conclusions	67
7.1	Discussions	67
7.1.1	Relation Extraction	67
7.1.2	Graph Ranking Model	68
7.1.3	Term Category Dependency	68
7.1.4	Retained Issues	69
7.2	Conclusions	70
	Bibliography	71

LIST OF FIGURES

Figure Number	Page
2.1 Knowledge based semantic interpreter presented by Gabrilovich & Markovitch (2007, 2009).	15
2.2 Out-link categories of the concepts “Machine Learning”, “Data Mining”, and “Computer Network”, presented by Wang & Domeniconi (2008).	16
3.1 The proposed framework of extracting relations	23
3.2 A linkage structure of a sentence derived from Stanford parser	24
5.1 The proposed term weighting framework based on category information	39
5.2 Classification results from OHSUMED corpus within 2^3 categories	51

LIST OF TABLES

Table Number	Page	
3.1	The example of relations extracted from the given sentence “Elephant garlic has roles in the prevention of cardiovascular disease. The “amod” means that the pair “elephant” and “garlic”, or the pair “cardiovascular” and “disease” has “ <i>noun compound modifier</i> ” relationships.	22
3.2	The example of built-in and associated relations extracted from the given sentence “ <i>Elephant garlic has roles in the prevention of cardiovascular disease</i> ”	26
3.3	Numbers of documents and categories from three experimental corpora	29
3.4	Numbers of built-in and associative relations of both training and testing sets from three different corpora	30
5.1	Numbers of documents and numbers of selected features (unique words with $DF \geq 3$) from the pre-processing step	47
5.2	Classification results from the Reuters-21578 corpus. The \star indicates that the precision column was reported by Joachims (1998)	49
5.3	Classification results on all corpora from the round-robin process. The table shows the results of the first 5 loops and of the final convergence step. The number in bracket indicates the position that convergence status is detected. The aggregated results are the outcome results of the framework	50
6.1	Statistic information of three corpora after pre-processing based on TF \times IDF method	55
6.2	Statistical information about TCO-relations with noun-phrase filter from three corpora after pre-processing	57
6.3	Statistic information of three corpora after pre-processing for RW \times IDF method. The features was selected as components of noun-phrases	57
6.4	Six term weighting methods implemented for the experiment	58
6.5	Classification judgement for each category i , where TP , FP , FN is denoted as <i>true positive</i> , <i>false positive</i> , <i>false negative</i>	60

6.6	The document frequency measure is used to reduce the large number of features. The table shows the statistic information of original features, and selected features on the experimental corpora used for PR×IDF and PR×TCD	62
6.7	Text classification results (micro-averaged F_1 measure) from TCD-based methods using the Linear SVM.	63
6.8	Classification results are detailed from the TCD framework, where $\delta = 0$ and the number of loops is selected as 5. The table shows results of the first 5 loops and of the aggregated results. The number in bracket indicates the iteration where the convergence was detected. The aggregated results are the outcome results of the framework . . .	63
6.9	Text classification results (micro-averaged F_1 measure) from IDF-based methods using the Linear SVM	64
6.10	Text classification results using SVM for six weighting schemas	65

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