Faculty of Engineering and Information Technology
University of Technology, Sydney

Representing Semantic Relatedness

A thesis submitted in partial fulfillment of
the requirements for the degree of
Master of Analytics by Research

by

Qianqian Chen

April 2016

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

_____

Acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisors Prof. Longbing Cao, Dr. Guandong Xu and Dr. Wei Liu for the continuous support of my study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

I also would like to appreciate my co-author Liang Hu for providing me with continuous support throughout my study and research. Without his professional guidance and persistent help, this thesis would not have been possible.

In addition, I thank my fellow lab mates in Advanced Analystics Institute: Jia Xu and Jingyu Shao for the stimulating discussions, for the hard days we were working together, and for all the fun we have had in the last three years.

Last but not the least, I would like to thank my family, for their unconditional support, both financially and emotionally throughout my whole study.

Qianqian Chen
December 2015 @ UTS

Contents

Certificate . . . . . **i**
Acknowledgment . . . . . **iii**
List of Figures . . . . . **ix**
List of Tables . . . . . **xi**
List of Publications . . . . . **xiii**
Abstract . . . . . **xv**

Chapter 1 Introduction **1**
1.1 Background 1
1.1.1 Non-iidness Learning 1
1.1.2 Representing Semantic 3
1.2 Research Issues 5
1.2.1 Capturing Coupled Pairwise Term Semantics 6
1.2.2 Measuring High Order Semantics by Hierarchical Tree Learning 7
1.2.3 Managing Natural Language Ambiguities 8
1.3 Research Contributions 9
1.4 Thesis Structure 10

Chapter 2 Literature Review **12**
2.1 Semantics 13
2.2 Corpus-based Semantic Representation 14
2.2.1 Term-based Semantic Representation 14
2.2.2 Topic-based Semantic Representation 29

2.2.3 Discussion and Conclusion 40
2.3 Lexical Resource-based Semantic Representation 41
2.3.1 Edge-based Semantic Representation 42
2.3.2 Information-theoretic Semantic Representation 43
2.3.3 Wikipedia-based Semantic Representation 50
2.3.4 Discussion and Conclusion 54
2.4 Summary 55

Chapter 3 Semantic Representation: Capturing Explicit and Implicit Content Couplings 57
3.1 Introduction 57
3.2 Term Pair Semantic Coupling Analysis 61
3.2.1 Semantic Intra-couplings within Term Pairs 62
3.2.2 Inter-couplings between Term Pairs 66
3.2.3 Semantic Couplings of Term Pairs 73
3.3 Coupled Document Analysis 75
3.4 Experiments and Evaluation 76
3.4.1 Experimental Settings 77
3.4.2 Inter-coupling Ordering 79
3.4.3 Tuning Parameter α 81
3.4.4 Experimental Results 82
3.5 Conclusions 83

Chapter 4 Semantic Representation Using Hierarchical Tree Augmented Naive Bayes 91
4.1 Introduction 91
4.2 Bayes Classification Methods 94
4.2.1 Naive Bayes Classifier 95
4.2.2 Tree Augmented Naive Bayes Classifier 97
4.3 Hierarchical Tree Learning 100
4.3.1 Feature Extraction 100
4.3.2 Feature Correlation 103

4.4 Hierarchical Structured TAN (HTAN) 107
4.4.1 The Construct-HTAN Procedure 107
4.4.2 Using HTAN as Text Classifiers 110
4.5 Conclusions 112

Chapter 5 Conclusions and Future Work 114
5.1 Conclusions 114
5.2 Future Work 115

Appendix A List of Symbols 117

Bibliography 120

List of Figures

1.1 The profile of work in this thesis 11

2.1 Cosine similarity in vector space model 19
2.2 An example of the inter-relation between terms 27
2.3 Graphical Representation of PLSA 32
2.4 Graphical Representation of LDA 35

3.1 An overview of term pair semantic coupling analysis 61
3.2 The Term Pair Frequency Graph 66
3.3 The Intra-coupling Graph 67
3.4 The Inter-coupling Graph 68
3.5 Tuning of α (1) 87
3.6 Tuning of α (2) 88

4.1 The structure of the naive Bayes network 95
4.2 A simple tree augmented naive Bayes structure 99
4.3 Schematic diagram of feature extraction procedure in HTAN . 101
4.4 The structure of the hierarchical tree augmented naive Bayes network 109

List of Tables

2.1 The Summary of Semantic Representations 56

3.1 Characteristics of Data Sets 78
3.2 The Impact of Inter-coupling Ordering Using HAC with Complete Linkage 85
3.3 The Impact of Inter-coupling Ordering Using HAC with Average Linkage 86
3.4 Results of Different Model Using HAC with Complete Linkage 89
3.5 Results of Different Model Using HAC With Average Linkage 90

List of Publications

Papers Published

- **Qianqian Chen**, Liang Hu, Jia Xu, Wei Liu, Longbing Cao (2015), Document Similarity Analysis via Involving both Explicit and Implicit Semantic Couplings. *in* 'Proceedings of the 2015 IEEE International Conference on Data Science and Advanced Analysis (**DSAA2015**)'.

Papers to be Submitted/Under Review

- **Qianqian Chen**, Liang Hu, Jia Xu, Longbing Cao (2015) Document Representation Using Hierarchical Tree Augmented Naive Bayes, to be submitted.

- **Qianqian Chen**, Liang Hu, Jia Xu, Longbing Cao (2015) A Novel Document Similarity Representation: Capturing Explicit and Implicit Content Couplings, to be submitted.

Abstract

To do text mining, the first question we must address is how to represent documents. The way a document is organised reflects certain explicit and implicit semantic and syntactical coupling relationships which are embedded in its contents. The effective capturing of such content couplings is thereby crucial for a genuine understanding of text representations. It has also led to the recent interest in document similarity analysis, including semantic relatedness, content coverage, word networking, and term-term couplings.

Document similarity analysis has become increasingly relevant since roughly 80% of big data is unstructured. Accordingly, semantic relatedness has generated much interest owing to its ability to extract coupling relationships between terms (words or phrases). Existing work has focused more on explicit couplings and this is reflected in the models that have been built.

In order to address the research limitations and challenges associated with document similarity analysis, this thesis proposes a semantic coupling similarity measure and the hierarchical tree learning model to fully enrich the semantics within terms and documents, and represent documents based on the comprehensive couplings of term pairs. In contrast to previous work, the models proposed can deal with unstructured data and terms that are coupled for various reasons, thereby addressing natural language ambiguity problems.

Chapter 3 explores the semantic couplings of pairwise terms by involving three types of coupling relationships: (1) intra-term pair couplings, reflecting the explicit relatedness within term pairs that is represented by the relation

strength over probabilistic distribution of terms across document collection; (2) the inter-term pair couplings, capturing the implicit relatedness between term pairs by considering the relation strength of their interactions with other term pairs on all possible paths via a graph-based representation of term couplings; and finally, (3) semantic coupling similarity, which effectively combine the intra- and inter-term couplings. The corresponding term semantic similarity measures are then defined to capture such couplings for the purposes of analysing term and document similarity. This approach effectively addresses both synonymy (many words per sense) and polysemy (many senses per word) in a graphical representation, two areas that have up until now been overlooked by previous models.

Chapter 4 constructs a hierarchical tree-like structure to extract highly correlated terms in a layerwise fashion and to prune weak correlations in order to maintain efficiency. In keeping with the hierarchical tree-like structure, a hierarchical tree learning method is proposed. The main contributions of our work lie in three areas: (1) the hierarchical tree-like structure featuring hierarchical feature extraction and correlation computation procedures whereby highly correlated terms are merged into sets, and these are associated with more complete semantic information; (2) each layer is a maximal weighted spanning tree to prune weak feature correlations; (3) the hierarchical tree-like structure can be applied to both supervised and unsupervised learning approaches. In this thesis, the tree is associated with Tree Augmented Naive Bayes (TAN) as the Hierarchical Tree Augmented Naive Bayes (HTAN).

All of these models can be applied in the text mining tasks, including document clustering and text classification. The performance of the semantic coupling similarity measure is compared with typical document representation models on various benchmark data sets in terms of document clustering and classification evaluation metrics. These models provide insightful knowledge to organise and retrieve documents.