# Hierarchical Classification of Web Pages

by

**Wang Yi**

**Master Degree of Software Engineering**

**2009**




**Faculty of Science and Technology**
**University of Macau**

# Table of Contents

**List of Figures** ..... iv

**List of Tables** ..... v

**List of Abbreviations** ..... vi

**Acknowledgments** ..... vii

**Chapter 1: Introduction** ..... 1

1.1 Background ..... 1

1.2 Literature Review ..... 4

1.2.1 Hierarchical Classification ..... 4

1.2.2 Incremental Learning ..... 5

1.2.3 Multi-label Classification ..... 7

1.3 Motivation and Objective ..... 8

**Chapter 2: State of Art in Classification** ..... 11

2.1 Preprocessing of Text and Web Pages ..... 11

2.1.1 Term Extraction ..... 11

2.1.2 Term Weighting ..... 12

2.1.3 Dimensionality Reduction ..... 13

2.2 Support Vector Machines ..... 15

2.3 Other Classification Methods ..... 18

2.3.1 k-Nearest Neighbor Classifier ..... 18

2.3.2 Naïve Bayes Classifier ..... 19

2.3.3 Decision Tree Classifier ..... 20

2.4 Multi-label Classification using Classifiers ..... 22
2.4.1 One-to-rest Classifiers ..... 22
2.4.2 Pair-wise Classifiers ..... 24
2.4.3 Error Correcting Codes ..... 25
2.5 Performance Measures ..... 26
2.5.1 Precision, Recall and F~1 ..... 26
2.5.2 Loss function--$L_H$ ..... 27
**Chapter 3: Methodology of Hierarchical Classification ..... 30**
3.1 Basic Model ..... 30
3.2 Imbalanced Data Problem ..... 32
3.2.1 Problem Definition ..... 32
3.2.2 Problem Analysis ..... 33
3.2.2 Strategy -- Hierarchical Classification using Voting ..... 35
3.3 Incremental Learning with SVM ..... 37
3.3.1 Incremental Training ..... 37
3.3.2 Problem Analysis ..... 39
3.3.3 A New Incremental Learning Algorithm ..... 42
**Chapter 4: Experiments and Results ..... 44**
4.1 Datasets ..... 44
4.2 Document Indexing ..... 46
4.3 Experiments within VOTEM Algorithm ..... 47
4.3.1 Experiment results on *20newsgroups* ..... 47
4.3.2 Experiment results on *alibaba.com* web-page ..... 48
4.4 Experiments within *m-sv-inc* Algorithm ..... 49
4.4.1 Experiment results on *20newsgroups* ..... 49
4.4.2 Experiment results on *alibaba.com* web-page ..... 51
**Chapter 5: Conclusion & Further Work ..... 53**
**Bibliography ..... 55**

**APPENDIX A: Stopwords List** ..... **61**

**APPENDIX B: Categories Hierarchy -- *alibaba.com* Dataset** ..... **62**