

Automatic Hidden-Web Database Classification

by

Jingbai Zhang

Software Engineering Master Degree Program

2007



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

TABLE OF CONTENTS

LIST OF FIGURES	III
LIST OF TABLES	IV
LIST OF ABBREVIATIONS	V
CHAPTER 1: INTRODUCTION.....	1
1.1. BACKGROUND	1
1.2. MOTIVATION AND OBJECTIVE.....	4
1.3. ORGANIZATION OF THE THESIS.....	6
CHAPTER 2: RELATED WORK.....	7
2.1. SEARCH ENGINE	7
2.2. PARSING FORM.....	9
2.3. CONTENT EXTRACTION FROM HTML PAGES	10
2.4. WEB DOCUMENTS CLASSIFICATION	11
2.5. HIDDEN-WEB DATABASE CLASSIFICATION.....	13
CHAPTER 3: SYSTEM DESIGN AND DEVELOPMENT.....	15
3.1. SYSTEM OVERVIEW	17
3.2. THE CRAWLER.....	20
3.3. HTML PREPROCESSOR	22
3.4. FORM PARSER.....	24
3.4.1. Heuristics for parsing HTML forms	26
3.4.2. Form Discarding Policies.....	28
3.4.3. Label Assignment.....	29
3.5. CATEGORIES FEATURE EXTRACTION	31
3.5.1. Classification Schemes for Hidden-Web Databases	31
3.5.2. Category Features Extraction.....	32
3.5.2.1. Procedure for Category Features Extraction.....	33
3.5.2.2. Identifying Data-rich Section from HTML Web Pages	35
3.5.2.3. Pre-Processing Terms.....	36

3.5.2.4. Computing Terms Weight	37
3.6. FORM FILLING-OUT AGENT	40
3.6.1. Type of HW-DB Interface.....	40
3.6.2. Type of Form Element	42
3.6.3. Generation of Queries	42
3.6.4. Submitting Forms Filled-out.....	43
3.7. RESPONSE PAGE ANALYZER.....	46
3.7.1. Determining the Error Result Page	46
3.7.2. Computing the Number of Results Documents	47
3.7.3. Extracting Knowledge of HW-DB.....	48
3.8. NEIGHBOR PAGE ANALYZER	51
3.9. HW-DB CLASSIFIER.....	56
CHAPTER 4: EXPERIMENTS	59
4.1. EXPERIMENTAL SETTING.....	59
4.1.1. Data Collections.....	59
4.1.2. Parameters Setting	60
4.1.3. Techniques for Evaluation	60
4.2. EXPERIMENT RESULTS	63
4.2.1. Determining the Number of the Features for Form Filling-out	63
4.2.2. Evaluating Results over Different Classification Approaches.....	64
CHAPTER 5: CONCLUSIONS AND FUTURE WORK.....	66
BIBLIOGRAPHY	68
APPENDIX A: HIDDEN-WEB DATABASES FOR EXPERIMENTS	76