Business Information Extraction from Web

by

Lam Man I

Master of Science in E-Commerce Technology

2008



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 Research Motivation and Objective 2
1.3 Organization of the Thesis 4
Chapter 2: Literature Review 5
2.1 Text Information Source 5
2.2 Critical Challanges 6
2.3 Overview of Existing System 7
Chapter 3: Web Extraction Based on HTML Analysing Approach 10
3.1 Approach Introduction 10
3.2 System Overview 10
3.3 "Yellow Page" Pages Collection Process 11
3.4 Extraction Process 14
3.4.1 Extraction Rule Generation Example 14
3.5 Evaluation AND further improvement 18
Chapter 4: Web Extraction Based on XSLT Pattern Template Approach 20
4.1 Approach Introduction 20
4.2 System Overview 20
4.3 Transformation Process 21
4.4 Extraction Process 22
4.5 Evaluation and Further Improvement 27
Chapter 5: Web Extraction Based on Pattern Classification Approach 32
5.1 Approach Introduction 32
5.2 System Overview 32
5.3 Pages Collection 34

5.4 Page Transformation....36
5.4.1 SGMLReader Transformation....36
5.4.2 Tidy Transformation....36
5. 5 Human Training & Extraction Information Gathering....36
5.6 Information Extraction Processor....39
5.6.1 Methodology....39
5.6.2 Introduction to Regular Expression....39
5.6.3 Regular Expression Rule Generation....40
5.6.4 Target Information Extraction....41
5.7 Evaluation and Further Improvement....42
Chapter 6: Web Extraction Based on Extraction Field Classification Approach....44
6.1 Approach Introduction....44
6.2 System Overview....44
6.3 Patterns and Rules of the Information....46
6.4 Information Extraction Processing....48
6.5 Methodology for Static Structure Information....48
6.5.1 Rule Extraction....49
6.5.2 Target Information Extraction....50
6.6 Methodology for Non-Static Structure Information....50
6.7 Enhancement....54
6.8 Evaluation and Further Improvement....54
Chapter 7: Conclusions and Future Work....57
BIBLIOGRAPHY....60