Web Image Clustering and Retrieval

by

U Leong Hou

E-Commerce Technology Master Degree Program

2005

Faculty of Science and Technology
University of Macau

TABLE OF CONTENTS

CHAPTER 1: Introduction 1

1.1. Background 1

1.1.1. Information Retrieval 1

1.1.2. Web Image Retrieval 3

1.1.3. Data Mining 5

1.2. Research Motivation and Objectives 7

1.3. Overview 9

CHAPTER 2: Related work 10

2.1. Web Image Retrieval 10

2.2. Data Mining Technology 12

2.2.1. Association Analysis 12

2.2.2. Clustering Technology 13

2.2.3. The Vector Space Model and Document Clustering 16

2.2.4. Some existed researches on document clustering model 18

2.3. Web Mining Segmentation 20

2.4. Web Mining Technology 23

2.4.1. Hypertext-Induced Topic Selection 24

2.4.2. Page Rank 24

CHAPTER 3: Web Image Retrieval 26

3.1. System Overview 26

3.2. The Crawler and HTML Preprocessor 27

3.3. The Semantic Extractor 30

3.3.1. The problem 30

3.3.2. Text Fragmentation 31

3.3.3. Distance Adjustment of Semantic Blocks 33

3.3.4. Term Weight Calculation 36

3.4. The Indexer 38

3.5. The Knowledge Base 40

3.6. The Query Engine 42
3.7. Evaluation of the Results 44
3.8. Conclusions and Future Work 49
CHAPTER 4: Clustering By Text 50
4.1. System Overview 50
4.2. Semantic Extractor 52
4.3. Building Term Semantic Network 53
4.4. Web Image Clustering 55
4.4.1. Term Clustering 55
4.4.2. Extracting Semantic Significances of Terms 58
4.4.3. Assigning Web Images to Corresponding Clusters 60
4.5. Experiments 61
4.6. Conclusions and Future Work 64
CHAPTER 5: Web Page Segamentation And Link Analysis 65
5.1. System Overview 65
5.2. Semantic Similarity Between Two Text Blocks 67
5.2.1. Semantic Similarity Between Two Text Blocks in General Document 67
5.2.2. Semantic Coherence of Two Blocks in HTML 68
5.3. Web Page Segmentation 72
5.3.1. Reducing the DOM Tree Size 72
5.3.2. Web Page Segmentation 74
5.3.3. Evaluations for Different Segmentation Methods 80
5.4. Block Based Link Analysis 87
5.4.1. The Importance for Web Page Block 87
5.4.2. Block Based Link Analysis 88
5.4.3. Retrieval Performance 90
5.5. Conclusions and Future Work 95
CHAPTER 6: Framework Conclusion 97
bibliography 99
APPENDIX A: Sample Queries Used In This Work 105

APPENDIX B: Summary Of Statistic Evaluation Result.....106