

SEMANTIC-ENHANCED WEB-PAGE RECOMMENDER SYSTEMS

A dissertation submitted for the degree of
Doctor of Philosophy in Computing Sciences

By

Thi Thanh Sang Nguyen

Faculty of Engineering and Information Technology
UNIVERSITY OF TECHNOLOGY, SYDNEY

Australia
December, 2012

CERTIFICATE OF 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 Student

Production Note:
Signature removed prior to publication.

ACKNOWLEDGEMENTS

Undertaking and completing a task like this would not have been possible without the encouragement and support of many individuals.

First and foremost, I would like to pay my respects and thanks to my supervisors, Dr. Haiyan (Helen) Lu and Prof. Jie Lu, for their technical ideas and constructive criticism which has significantly contributed to the success of this thesis. Their professional advice has been a great asset to have during my moments of confusion and I truly acknowledge that.

I also acknowledge the great help and cooperation that I received from the staff of UTS library and the School of Software, Faculty of Engineering and Information Technology (FEIT), as well as my colleagues in the Decision Systems and e-Service Intelligence (DeSI) laboratory within the Centre for Quantum Computation and Intelligent Systems (QCIS).

I am grateful to my dear friend, Dr. Tich Phuoc Tran, for his useful advice and friendly support during my PhD candidature.

Last but not least, I would also like to thank my family, who gave me the opportunity for my education and supported me in every direction. Their love and care enabled me to ease my mind and soul so that I could concentrate on this thesis.

Sydney, Australia, December, 2012.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
TABLE OF CONTENTS.....	iii
LIST OF FIGURES.....	viii
LIST OF TABLES.....	x
LIST OF ALGORITHMS.....	xii
ABSTRACT	xiii
Chapter 1. INTRODUCTION.....	1
1.1. Background	1
1.2. Research Questions and Objectives	4
1.3. Research Significance and Contributions.....	7
1.4. Organization of the Thesis	10
1.5. Publications Related to the Thesis.....	14
Chapter 2. LITERATURE REVIEW	15
2.1. Introduction	15
2.2. Web Mining.....	16
2.2.1. Web Content Mining.....	16
2.2.2. Web Structure Mining.....	16
2.2.3. Web Usage Mining	17
2.3. Ontology.....	20
2.3.1. Definitions of an Ontology.....	20
2.3.2. Roles of Ontology.....	23
2.3.3. Ontology Languages.....	24
2.3.4. Variants of Ontologies.....	25
2.3.5. Examples of Ontologies.....	27
2.3.6. Ontology Construction	28
2.3.7. Ontology Learning.....	30
2.3.8. Ontology Reasoning	32
2.3.9. Ontology Evaluation.....	33
2.4. Recommender Systems	35
2.4.1. Basic Types of Recommender Systems	36

2.4.2.	Semantic Recommender Systems.....	38
2.4.3.	Web-page Recommender Systems.....	39
2.4.4.	Recommender System Evaluation	41
2.5.	Summary	43
Chapter 3.	SEQUENTIAL PATTERN MINING FOR WEB USAGE	44
3.1.	Introduction	44
3.2.	The Web Usage Data as the Source of Mining.....	45
3.3.	Sequential Pattern Mining	47
3.3.1.	Sequential Pattern Mining Algorithms.....	47
3.3.2.	Pre-order Linked Web Access Pattern Tree Mining.....	51
3.3.3.	Conditional Sequence Mining.....	51
3.3.4.	Comparison.....	52
3.4.	Web-page Recommendation using Tree-based Mining Techniques.....	53
3.4.1.	Web-page Recommender System Architecture	53
3.4.2.	Sequential Pattern Mining Component	55
3.4.3.	Frequent Web Access Pattern Tree Construction Component	56
3.4.4.	Recommendation Rule Generation Component	57
3.5.	Experiments with the Two Chosen Sequential Pattern Mining Algorithms.....	58
3.5.1.	Evaluation of Sequential Pattern Mining Algorithms.....	58
3.5.2.	Training and Testing Datasets.....	60
3.5.3.	Experimental Results and Analysis.....	61
3.6.	Remarks	70
3.7.	Summary	71
Chapter 4.	DOMAIN ONTOLOGY MODELLING FOR WEB-PAGE RECOMMENDATION	72
4.1.	Introduction	72
4.2.	Domain Ontology Model of a Website	73
4.2.1.	Assumption.....	73
4.2.2.	Domain Ontology Model of a Website.....	75
4.3.	New Method for Domain Ontology Modelling of a Website	77
4.4.	Reasoning Algorithms for the Domain Ontology Model of Web-pages	83
4.5.	Case Study: Development of a Domain Ontology of the Microsoft Website	84

4.5.1.	Requirements Analysis	84
4.5.2.	Conceptualization	85
4.5.3.	Implementation	88
4.6.	Evaluation and Discussion.....	93
4.6.1.	Evaluation Method.....	94
4.6.2.	Evaluation of Experimental Results, and Discussions	94
4.7.	Summary	96
Chapter 5.	SEMANTIC NETWORK MODELLING FOR WEB-PAGE RECOMMENDATION	98
5.1.	Introduction	98
5.2.	Preparation for Semantic Network Construction.....	99
5.3.	A New Method to Automatically Construct a Semantic Network of Web-pages	101
5.4.	Modelling of Semantic Network of Web-pages	103
5.4.1.	Term Extraction Algorithm	103
5.4.2.	Construction of Semantic Network of Web-pages	105
5.4.3.	Reasoning Algorithms for the Semantic Network of Web-pages	110
5.4.4.	An Experimental Example of Semantic Network Construction	117
5.5.	Experimental Evaluation	121
5.5.1.	Evaluation Measures.....	121
5.5.2.	Experiment Results	122
5.6.	Remarks	123
5.7.	Summary	124
Chapter 6.	CONCEPT NAVIGATION MODELS FOR PREDICTION	126
6.1.	Introduction	126
6.2.	A Web Usage Knowledge Representation Model of a Website for Web-page Recommendation.....	127
6.2.1.	A Web-page Navigation Model.....	127
6.2.2.	Schema of Web-page Navigation Model.....	131
6.2.3.	Automatic Construction of Web-page Navigation Model.....	132
6.2.4.	Reasoning Algorithms for the Web-page Navigation Model.....	133
6.3.	A Semantic Web Usage Knowledge Representation Model for Web-Page Recommendation.....	135
6.3.1.	A Domain Term Navigation Model.....	137

6.3.2.	Schema of Domain Term Navigation Model.....	139
6.3.3.	Automatic Construction of Domain Term Navigation Model.....	140
6.3.4.	Reasoning Algorithms for the Domain Term Navigation Model.....	141
6.4.	An Example of Using the Proposed Navigation Models	144
6.5.	Summary	149
Chapter 7.	A SEMANTIC-ENHANCED WEB-PAGE RECOMMENDER SYSTEM FRAMEWORK	151
7.1.	Introduction	151
7.2.	Framework	152
7.3.	Web-page Recommendation Strategies	157
7.4.	Experimental Evaluation	174
7.4.1.	Performance Evaluation	175
7.4.2.	Experimental Method	176
7.4.3.	Experiments with Public Datasets.....	177
7.4.4.	Experiments with Real World Dataset	195
7.5.	Discussions	201
7.6.	Summary	202
Chapter 8.	A SEMANTIC-ENHANCED WEB-PAGE RECOMMENDER SYSTEM PROTOTYPE	203
8.1.	Introduction	203
8.2.	System Overview	204
8.2.1.	System Statement and Scope	204
8.2.2.	Major Constraints	205
8.3.	System Architecture	206
8.3.1.	System Architecture Model.....	206
8.3.2.	Sub-System Overview	208
8.4.	Structural Modelling	212
8.4.1.	Pre-processing Sub-System	212
8.4.2.	Web Usage Mining Sub-System.....	214
8.4.3.	Domain Knowledge Construction Sub-System	217
8.4.4.	Prediction Model Sub-System	223
8.4.5.	Recommendation Engine Sub-System	229
8.5.	Operation	231

8.5.1.	Pre-processing	231
8.5.2.	Web Usage Mining	234
8.5.3.	Domain Knowledge Construction	234
8.5.4.	Prediction Model	235
8.5.5.	Web-page Recommendation.....	237
8.6.	Interface Description	238
8.6.1.	Back-end.....	239
8.6.2.	Front-end.....	244
8.7.	Summary	245
Chapter 9.	CONCLUSIONS AND FUTURE RESEARCH	247
9.1.	Conclusions	247
9.2.	Future Research.....	251
ABBREVIATIONS		254
BIBLIOGRAPHY		255

LIST OF FIGURES

Figure 1-1: The overall structure of the thesis 13

Figure 2-1: Topic map..... 15

Figure 2-2: Excerpt of a domain ontology for personalized e-learning in educational systems (Boyce & Pahl 2007) 22

Figure 2-3: Architecture for learning ontologies for the Semantic Web (Maedche 2002) 31

Figure 3-1: Sample Web log from website <http://www.usask.ca/> 46

Figure 3-2: Web-page recommender system architecture 54

Figure 3-3: Sequence length of (a) NASA, (b) Sask, and (c) Cezeife 61

Figure 3-4: Execution times of the two mining algorithms with different supports 63

Figure 3-5: The number of frequent patterns obtained from the two mining algorithms with different supports 65

Figure 3-6: Performance of Web-page recommendation based on PLWAP-Mine vs. CS-Mine..... 70

Figure 4-1: Sample Web document 74

Figure 4-2: The flow diagram for domain ontology modelling 77

Figure 4-3: Web-page mapping 81

Figure 4-4: Domain ontology schema of the MS website 86

Figure 4-5: Specification of object properties in the domain ontology of the MS website 89

Figure 4-6: The part of the domain ontology: *Product* instances 90

Figure 4-7: The part of the domain ontology: *Product* instance - *Word* 93

Figure 5-1: Flow diagram for automatic construction of a semantic network of Web-pages 101

Figure 5-2: Process of collecting the accessed Web-pages 102

Figure 5-3: Process of extracting domain terms..... 102

Figure 5-4: Graphical representation of TermNetWP 107

Figure 5-5: The schema of TermNetWP 107

Figure 5-6: TermNetWP population: A set of titles => the corresponding term instances and relations in the semantic network of Web-pages..... 120

Figure 6-1: The first-order Web-page navigation model with respect to the patterns given in Table 6-1 131

Figure 6-2: The schema of Web-page navigation model 132

Figure 6-3: Frequently viewed domain term pattern discovery..... 136

Figure 6-4: Building a domain term navigation network 137

Figure 6-5: The schema of domain term navigation model 139

Figure 6-6: A sample set of frequent Web access patterns 144

Figure 6-7: A sample 1st-order WPNavNet 145

Figure 6-8: A sample 2nd-order WPNavNet 145

Figure 6-9: A sample set of Web-pages and titles 147

Figure 6-10: A sample TermNavNet..... 148

Figure 7-1: Framework of the semantic-enhanced Web-page recommender system 154

Figure 7-2: Results for experimental Cases 1-5 184

Figure 7-3: Results for experimental Cases 4-7 185

Figure 7-4: Results for experimental Cases 1, 2, 3, 6, and 7..... 186

Figure 7-5: Results for experimental Cases 6-9 and 12,13 188

Figure 7-6: Results for experimental Cases 6, 7, 10, 11, 14, and 15 189

Figure 7-7: Results for experimental Cases 1, 6, 8, 10, 12, and 14 191

Figure 7-8: Results for experimental Cases 1, 7, 9, 11, 13, and 15 192

Figure 7-9: The medians of Precisions and Satisfactions of Cases 1-3, 6-15 194

Figure 7-10: Experiment 1: the resulting satisfactions of R.WP.AutoTopic.1st.4 and
R.WP.AutoTopic.1st.5 vs R.PLWAP 198

Figure 7-11: Experiment 2: the resulting satisfactions of R.WP.AutoTopic.1st.4 and
R.WP.AutoTopic.1st.5 vs R.PLWAP 199

Figure 7-12: Experiment 3: the resulting satisfactions of R.WP.AutoTopic.1st.4 and
R.WP.AutoTopic.1st.5 vs R.PLWAP 200

Figure 8-1: Semantic-enhanced Web-page recommender system architecture 207

Figure 8-2: ERD storing Web access information of WebCleaner 213

Figure 8-3: The WUM package 215

Figure 8-4: The DOC package..... 218

Figure 8-5: The SNC package 221

Figure 8-6: The domain term navigation model package 225

Figure 8-7: The RE package..... 229

Figure 8-8: Sample Web log of website *handbook.uts.edu.au* 231

Figure 8-9: Sample dataset obtained from the Web log of website *handbook.uts.edu.au* 233

Figure 8-10: Sample TermNetWP in the Protégé interface..... 235

Figure 8-11: Sample TermNavNet in the Protégé interface 236

Figure 8-12: Web-page recommendation results of page *utshb_719* 237

Figure 8-13: Web-page recommendation results of page *utshb_719* and *utshb_127* 238

Figure 8-14: Main frame of the semantic-enhanced Web-page recommender system..... 239

Figure 8-15: Pre-processing frame..... 240

Figure 8-16: Web usage mining frame 241

Figure 8-17: Semantic network construction frame 242

Figure 8-18: Conceptual prediction model frame..... 243

Figure 8-19: Recommendation engine frame..... 244

Figure 8-20: Web browser frame..... 245

LIST OF TABLES

Table 3-1: The performance of sequential pattern mining algorithms..... 53

Table 3-2: Statistic of the three real world Web access sequence datasets 60

Table 4-1: Keyword expressions 80

Table 4-2: A Sample MS Web dataset..... 85

Table 4-3: Domain concepts and corresponding domain terms 85

Table 4-4: Mapping some Web-pages to some domain terms based on the specified keyword strings 91

Table 4-5: Evaluation of the domain ontology of the MS website 95

Table 5-1: Sample of extracted domain term sequences 105

Table 5-2: TermNetWP evaluation..... 122

Table 6-1: A set of Web access patterns 130

Table 6-2: Web-page prediction cases 148

Table 7-1: Procedure of Web-page recommendation generation 156

Table 7-2: Web-page recommendation strategies 160

Table 7-3: Description logics notation of strategies R.WP.ManTopic.1st.1 and R.WP.ManTopic.2nd.1 166

Table 7-4: Description logics notation of strategies R.WP.AutoTopic.1st.1 and R.WP.AutoTopic.2nd.1 167

Table 7-5: Description logics notation of strategies R.WP.AutoTopic.1st.2 and R.WP.AutoTopic.2nd.2 169

Table 7-6: Description logics notation of strategies R.WP.AutoTopic.1st.3 and R.WP.AutoTopic.2nd.3 170

Table 7-7: Description logics notation of strategies R.WP.AutoTopic.1st.4 and R.WP.AutoTopic.2nd.4 172

Table 7-8: Description logics notation of strategies R.WP.AutoTopic.1st.5 and R.WP.AutoTopic.2nd.5 173

Table 7-9: Experimental cases 179

Table 7-10: Comparisons of experimental results 181

Table 8-1: Pre-processing 209

Table 8-2: Web usage mining 209

Table 8-3: Domain knowledge construction..... 210

Table 8-4: Prediction model 210

Table 8-5: Recommendation engine 211

Table 8-6: Web browser 212

Table 8-7: Tables in the data warehouse of WebCleaner 214

Table 8-8: The WUM package..... 214

Table 8-9: Data structure of class Node in the WUM package..... 215

Table 8-10: Data structure of class LinkHeader in the WUM package..... 215

Table 8-11: Data structure of class PLWAP in the WUM package 216

Table 8-12: The DOC package	217
Table 8-13: Data structure of class Page in the DOC package	218
Table 8-14: Data structure of class OntoPageAdd in the DOC package.....	219
Table 8-15: Data structure of class Reasoner in the DOC package.....	219
Table 8-16: The SNC package	220
Table 8-17: Data structure of class Page in the SNC package	221
Table 8-18: Data structure of class Instance in the SNC package.....	221
Table 8-19: Data structure of class OutLink in the SNC package.....	222
Table 8-20: Data structure of class CollocationMap in the SNC package	222
Table 8-21: Data structure of class Reasoner in the SNC package	223
Table 8-22: The domain term navigation model package.....	224
Table 8-23: Data structure of class Node in the domain term navigation model package	226
Table 8-24: Data structure of class InLink in the domain term navigation model package	226
Table 8-25: Data structure of class OutLink in the domain term navigation model package	227
Table 8-26: Data structure of class NavModel in the domain term navigation model package	227
Table 8-27: Data structure of class Reasoner in the domain term navigation model package.....	228
Table 8-28: The Web-page navigation model package.....	229
Table 8-29: The RE package.....	230
Table 8-30: Sample data in Table <i>UserDimTbl</i>	232
Table 8-31: Sample data in Table <i>ProtocolDimTbl</i>	232
Table 8-32: Sample data in Table <i>PathDimTbl</i>	232
Table 8-33: Sample data in Table <i>LogFactTbl</i>	232
Table 8-34: A sample collection of the titles and URLs of accessed Web-pages.....	233
Table 8-35: A sample FWAP discovered from a sample WAS with MinSup = 0.6 %.....	234

LIST OF ALGORITHMS

Algorithm 3-1: PLWAP-Mine..... 56

Algorithm 3-2: CS-Mine 56

Algorithm 3-3: FWAP-tree construction..... 57

Algorithm 3-4: Recommendation rule generation..... 57

Algorithm 3-5: Performance evaluation..... 59

Algorithm 4-1: Mapping Web-pages to domain terms using the keyword expressions..... 82

Algorithm 4-2: Query about domain terms of a given Web-page 83

Algorithm 4-3: Query about pages mapped to a given domain term..... 84

Algorithm 5-1: Term extraction 104

Algorithm 5-2: Automatically constructing a TermNetWP..... 109

Algorithm 5-3: Query about domain terms of a given Web-page 112

Algorithm 5-4: Query about pages mapped to a given domain term..... 113

Algorithm 5-5: Query about pages mapped to a given set of domain terms..... 114

Algorithm 5-6: Query about Web-pages mapped to a given set of domain terms, in which each domain term is assigned a prediction probability 116

Algorithm 5-7: Query about domain terms of a given Web-page assigned a prediction probability 117

Algorithm 6-1: WPNavNet construction 133

Algorithm 6-2: Query about next pages for a given current page and a given previous page..... 134

Algorithm 6-3: Query about next pages for a given current page..... 135

Algorithm 6-4: TermNavNet construction..... 140

Algorithm 6-5: Query about next domain terms for a given current domain term and a given previous domain term..... 141

Algorithm 6-6: Query about next domain terms for a given current domain term 142

Algorithm 6-7: Query about next domain terms and respective prediction probabilities for a given current domain term and a given previous domain term..... 143

Algorithm 6-8: Query about next domain terms and respective prediction probabilities for a given current domain term..... 143

Algorithm 7-1: Web-page recommendation without semantic enhancement..... 165

Algorithm 7-2: Web-page recommendation with semantic enhancement based on DomainOntoWP 166

Algorithm 7-3: Web-page recommendation with semantic enhancement based on TermNetWP 168

Algorithm 7-4: Web-page recommendation with semantic enhancement based on TermNetWP 169

Algorithm 7-5: Web-page recommendation with semantic enhancement based on TermNetWP 171

Algorithm 7-6: Web-page recommendation with semantic enhancement based on TermNetWP 172

Algorithm 7-7: Web-page recommendation with semantic enhancement based on TermNetWP 174

Algorithm 7-8: Performance evaluation..... 176

ABSTRACT

This thesis presents a new framework for a semantic-enhanced Web-page recommender (WPR) system, and a suite of enabling techniques which include semantic network models of domain knowledge and Web usage knowledge, querying techniques, and Web-page recommendation strategies. The framework enables the system to automatically discover and construct the domain and Web usage knowledge bases, and to generate effective Web-page recommendations. The main contributions of the framework are fourfold: (1) it effectively changes the fact that knowledge base construction must rely on human experts; (2) it enriches the pool of candidate Web-pages for effective Web-page recommendations by using semantic knowledge of both Web-pages and Web usage; (3) it thoroughly resolves the inconsistency problem facing contemporary WPR systems which heavily employ heterogeneous representations of knowledge bases. Knowledge bases in the system are consistently represented in a formal Web ontology language, namely OWL; and (4) it can generate effective Web-page recommendations based on a set of thoughtfully-designed recommendation strategies. A prototype of the semantic-enhanced WPR system is developed and presented, and the experimental comparisons with existing WPR approaches convincingly prove the significantly improved performance of WPR systems based on the framework and its enabling techniques.

Keywords: Web-page recommender systems, Web usage mining, domain knowledge modelling, knowledge representation, semantic network, semantic reasoning