


ZDRAVKO MARKOV
DANIEL T. LAROSE

 WILEY

DATA MINING THE WEB

UNCOVERING PATTERNS IN
WEB CONTENT, STRUCTURE,
AND USAGE

www

COMPANION WEB SITE AVAILABLE

Wiley Series on Methods and Applications in Data Mining • Daniel T. Larose, Series Editor

CONTENTS

PREFACE

xi

PART I

WEB STRUCTURE MINING

1	INFORMATION RETRIEVAL AND WEB SEARCH	3
	Web Challenges	3
	Web Search Engines	4
	Topic Directories	5
	Semantic Web	5
	Crawling the Web	6
	Web Basics	6
	Web Crawlers	7
	Indexing and Keyword Search	13
	Document Representation	15
	Implementation Considerations	19
	Relevance Ranking	20
	Advanced Text Search	28
	Using the HTML Structure in Keyword Search	30
	Evaluating Search Quality	32
	Similarity Search	36
	Cosine Similarity	36
	Jaccard Similarity	38
	Document Resemblance	41
	References	43
	Exercises	43
2	HYPERLINK-BASED RANKING	47
	Introduction	47
	Social Networks Analysis	48
	PageRank	50
	Authorities and Hubs	53
	Link-Based Similarity Search	55
	Enhanced Techniques for Page Ranking	56
	References	57
	Exercises	57

PART II**WEB CONTENT MINING**

3	CLUSTERING	61
	Introduction	61
	Hierarchical Agglomerative Clustering	63
	<i>k</i> -Means Clustering	69
	Probability-Based Clustering	73
	Finite Mixture Problem	74
	Classification Problem	76
	Clustering Problem	78
	Collaborative Filtering (Recommender Systems)	84
	References	86
	Exercises	86
4	EVALUATING CLUSTERING	89
	Approaches to Evaluating Clustering	89
	Similarity-Based Criterion Functions	90
	Probabilistic Criterion Functions	95
	MDL-Based Model and Feature Evaluation	100
	Minimum Description Length Principle	101
	MDL-Based Model Evaluation	102
	Feature Selection	105
	Classes-to-Clusters Evaluation	106
	Precision, Recall, and <i>F</i> -Measure	108
	Entropy	111
	References	112
	Exercises	112
5	CLASSIFICATION	115
	General Setting and Evaluation Techniques	115
	Nearest-Neighbor Algorithm	118
	Feature Selection	121
	Naive Bayes Algorithm	125
	Numerical Approaches	131
	Relational Learning	133
	References	137
	Exercises	138

PART III**WEB USAGE MINING**

6	INTRODUCTION TO WEB USAGE MINING	143
	Definition of Web Usage Mining	143
	Cross-Industry Standard Process for Data Mining	144
	Clickstream Analysis	147

Web Server Log Files	148
Remote Host Field	149
Date/Time Field	149
HTTP Request Field	149
Status Code Field	150
Transfer Volume (Bytes) Field	151
Common Log Format	151
Identification Field	151
Authuser Field	151
Extended Common Log Format	151
Referrer Field	152
User Agent Field	152
Example of a Web Log Record	152
Microsoft IIS Log Format	153
Auxiliary Information	154
References	154
Exercises	154
7 <i>PREPROCESSING FOR WEB USAGE MINING</i>	156
Need for Preprocessing the Data	156
Data Cleaning and Filtering	158
Page Extension Exploration and Filtering	161
De-Spidering the Web Log File	163
User Identification	164
Session Identification	167
Path Completion	170
Directories and the Basket Transformation	171
Further Data Preprocessing Steps	174
References	174
Exercises	174
8 <i>EXPLORATORY DATA ANALYSIS FOR WEB USAGE MINING</i>	177
Introduction	177
Number of Visit Actions	177
Session Duration	178
Relationship between Visit Actions and Session Duration	181
Average Time per Page	183
Duration for Individual Pages	185
References	188
Exercises	188
9 <i>MODELING FOR WEB USAGE MINING: CLUSTERING, ASSOCIATION, AND CLASSIFICATION</i>	191
Introduction	191
Modeling Methodology	192
Definition of Clustering	193
The BIRCH Clustering Algorithm	194
Affinity Analysis and the A Priori Algorithm	197

X CONTENTS

Discretizing the Numerical Variables: Binning	199
Applying the A Priori Algorithm to the CCSU Web Log Data	201
Classification and Regression Trees	204
The C4.5 Algorithm	208
References	210
Exercises	211

INDEX