CONTENTS

PREFACE xiv

LIST OF FIGURES xvii

CHAPTER 1 INTRODUCTION 1

1.1 Brief Summary of Chapters 2
1.2 Brief History of Hypertext and the Web 3
1.3 Brief History of Search Engines 6

CHAPTER 2 THE WEB AND THE PROBLEM OF SEARCH 9

2.1 Some Statistics 10
  2.1.1 Web Size Statistics 10
  2.1.2 Web Usage Statistics 15
2.2 Tabular Data Versus Web Data 18
2.3 Structure of the Web 20
  2.3.1 Bow-Tie Structure of the Web 21
  2.3.2 Small-World Structure of the Web 23
2.4 Information Seeking on the Web 24
  2.4.1 Direct Navigation 24
  2.4.2 Navigation within a Directory 25
  2.4.3 Navigation using a Search Engine 26
  2.4.4 Problems with Web Information Seeking 27
2.5 Informational, Navigational, and Transactional Queries 28
2.6 Comparing Web Search to Traditional Information Retrieval 29
  2.6.1 Recall and Precision 30
2.7 Local Site Search Versus Global Web Search 32
2.8 Difference Between Search and Navigation 34

CHAPTER 3 THE PROBLEM OF WEB NAVIGATION 38

3.1 Getting Lost in Hyperspace and the Navigation Problem 39
3.2 How Can the Machine Assist in User Search and Navigation 42
  3.2.1 The Potential Use of Machine Learning Algorithms 42
  3.2.2 The Naive Bayes Classifier for Categorizing Web Pages 43
3.3 Trails Should be First Class Objects 46
3.4 Enter Markov Chains and Two Interpretations of Its Probabilities 49
  3.4.1 Markov Chains and the Markov Property 49
  3.4.2 Markov Chains and the Probabilities of Following Links 50
  3.4.3 Markov Chains and the Relevance of Links 52
CONTENTS

3.5 Conflict Between Web Site Owner and Visitor  54
3.6 Conflict Between Semantics of Web Site and the Business Model  57

CHAPTER 4  SEARCHING THE WEB  60

4.1 Mechanics of a Typical Search  61
4.2 Search Engines as Information Gatekeepers of the Web  64
4.3 Search Engine Wars, is the Dust Settling?  68
  4.3.1 Competitor Number One: Google  69
  4.3.2 Competitor Number Two: Yahoo  70
  4.3.3 Competitor Number Three: Bing  70
  4.3.4 Other Competitors  72
4.4 Statistics from Studies of Search Engine Query Logs  73
  4.4.1 Search Engine Query Logs  73
  4.4.2 Search Engine Query Syntax  75
  4.4.3 The Most Popular Search Keywords  77
4.5 Architecture of a Search Engine  78
  4.5.1 The Search Index  79
  4.5.2 The Query Engine  80
  4.5.3 The Search Interface  81
4.6 Crawling the Web  81
  4.6.1 Crawling Algorithms  82
  4.6.2 Refreshing Web Pages  84
  4.6.3 The Robots Exclusion Protocol  84
  4.6.4 Spider Traps  85
4.7 What Does it Take to Deliver a Global Search Service?  85

CHAPTER 5  HOW DOES A SEARCH ENGINE WORK  91

5.1 Content Relevance  94
  5.1.1 Processing Web Pages  94
  5.1.2 Interpreting the Query  96
  5.1.3 Term Frequency  96
  5.1.4 Inverse Document Frequency  99
  5.1.5 Computing Keyword TF–IDF Values  100
  5.1.6 Caching Queries  102
  5.1.7 Phrase Matching  102
  5.1.8 Synonyms  102
  5.1.9 Link Text  103
  5.1.10 URL Analysis  104
  5.1.11 Date Last Updated  104
  5.1.12 HTML Structure Weighting  104
  5.1.13 Spell Checking  105
  5.1.14 Non-English Queries  106
  5.1.15 Home Page Detection  107
  5.1.16 Related Searches and Query Suggestions  107
5.2 Link-Based Metrics  108
  5.2.1 Referential and Informational Links  109
  5.2.2 Combining Link Analysis with Content Relevance  110
  5.2.3 Are Links the Currency of the Web?  110
CONTENTS

5.2.4 PageRank Explained 112
5.2.5 Online Computation of PageRank 116
5.2.6 Monte Carlo Methods in PageRank Computation 116
5.2.7 Hyperlink-Induced Topic Search 117
5.2.8 Stochastic Approach for Link-Structure Analysis 120
5.2.9 Counting Incoming Links 122
5.2.10 The Bias of PageRank against New Pages 123
5.2.11 PageRank within a Community 123
5.2.12 Influence of Weblogs on PageRank 124
5.2.13 Link Spam 125
5.2.14 Citation Analysis 127
5.2.15 The Wide Ranging Interest in PageRank 129

5.3 Popularity-Based Metrics 130
5.3.1 Direct Hit’s Popularity Metric 130
5.3.2 Document Space Modification 132
5.3.3 Using Query Log Data to Improve Search 132
5.3.4 Learning to Rank 133
5.3.5 BrowseRank 134

5.4 Evaluating Search Engines 136
5.4.1 Search Engine Awards 136
5.4.2 Evaluation Metrics 136
5.4.3 Performance Measures 138
5.4.4 Eye Tracking Studies 139
5.4.5 Test Collections 141
5.4.6 Inferring Ranking Algorithms 142

CHAPTER 6 DIFFERENT TYPES OF SEARCH ENGINES 148

6.1 Directories and Categorization of Web Content 150
6.2 Search Engine Advertising 152
6.2.1 Paid Inclusion 152
6.2.2 Banner Ads 153
6.2.3 Sponsored Search and Paid Placement 153
6.2.4 Behavioral Targeting 157
6.2.5 User Behavior 158
6.2.6 The Trade-Off between Bias and Demand 160
6.2.7 Sponsored Search Auctions 161
6.2.8 Pay per Action 165
6.2.9 Click Fraud and Other Forms of Advertising Fraud 166
6.3 Metasearch 168
6.3.1 Fusion Algorithms 169
6.3.2 Operational Metasearch Engines 170
6.3.3 Clustering Search Results 173
6.3.4 Classifying Search Results 175
6.4 Personalization 178
6.4.1 Personalization versus Customization 180
6.4.2 Personalized Results Tool 180
6.4.3 Privacy and Scalability 182
6.4.4 Relevance Feedback 182
6.4.5 Personalized PageRank 184
# CONTENTS

6.4.6 Outride’s Personalized Search 186

6.5 Question Answering (Q&A) on the Web 187
  6.5.1 Natural Language Annotations 188
  6.5.2 Factual Queries 190
  6.5.3 Open Domain Question Answering 191
  6.5.4 Semantic Headers 193

6.6 Image Search 194
  6.6.1 Text-Based Image Search 195
  6.6.2 Content-Based Image Search 196
  6.6.3 VisualRank 198
  6.6.4 CAPTCHA and reCAPTCHA 200
  6.6.5 Image Search for Finding Location-Based Information 200

6.7 Special Purpose Search Engines 201

<table>
<thead>
<tr>
<th>CHAPTER 7</th>
<th>NAVIGATING THE WEB</th>
<th>209</th>
</tr>
</thead>
</table>

7.1 Frustration in Web Browsing and Navigation 211
  7.1.1 HTML and Web Site Design 211
  7.1.2 Hyperlinks and Surfing 211
  7.1.3 Web Site Design and Usability 212

7.2 Navigation Tools 213
  7.2.1 The Basic Browser Tools 213
  7.2.2 The Back and Forward Buttons 214
  7.2.3 Search Engine Toolbars 215
  7.2.4 The Bookmarks Tool 216
  7.2.5 The History List 219
  7.2.6 Identifying Web Pages 219
  7.2.7 Breadcrumb Navigation 221
  7.2.8 Quicklinks 222
  7.2.9 Hypertext Orientation Tools 223
  7.2.10 Hypercard Programming Environment 224

7.3 Navigational Metrics 225
  7.3.1 The Potential Gain 226
  7.3.2 Structural Analysis of a Web Site 228
  7.3.3 Measuring the Usability of Web Sites 229

7.4 Web Data Mining 230
  7.4.1 Three Perspectives on Data Mining 230
  7.4.2 Measuring the Success of a Web Site 231
  7.4.3 Web Analytics 233
  7.4.4 E-Metrics 233
  7.4.5 Web Analytics Tools 234
  7.4.6 Weblog File Analyzers 235
  7.4.7 Identifying the Surfer 236
  7.4.8 Sessionizing 237
  7.4.9 Supplementary Analyses 237
  7.4.10 Markov Chain Model of Web Site Navigation 238
  7.4.11 Applications of Web Usage Mining 242
  7.4.12 Information Extraction 244

7.5 The Best Trail Algorithm 245
  7.5.1 Effective View Navigation 245
CONTENTS

7.5.2 Web Usage Mining for Personalization 246
7.5.3 Developing a Trail Engine 246

7.6 Visualization that Aids Navigation 252
7.6.1 How to Visualize Navigation Patterns 252
7.6.2 Overview Diagrams and Web Site Maps 253
7.6.3 Fisheye Views 255
7.6.4 Visualizing Trails within a Web Site 257
7.6.5 Visual Search Engines 258
7.6.6 Social Data Analysis 259
7.6.7 Mapping Cyberspace 262

7.7 Navigation in Virtual and Physical Spaces 262
7.7.1 Real-World Web Usage Mining 262
7.7.2 The Museum Experience Recorder 264
7.7.3 Navigating in the Real World 265

CHAPTER 8  THE MOBILE WEB  272

8.1 The Paradigm of Mobile Computing 273
8.1.1 Wireless Markup Language 274
8.1.2 The i-mode Service 275

8.2 Mobile Web Services 277
8.2.1 M-commerce 277
8.2.2 Delivery of Personalized News 278
8.2.3 Delivery of Learning Resources 281

8.3 Mobile Device Interfaces 282
8.3.1 Mobile Web Browsers 282
8.3.2 Information Seeking on Mobile Devices 284
8.3.3 Text Entry on Mobile Devices 284
8.3.4 Voice Recognition for Mobile Devices 286
8.3.5 Presenting Information on a Mobile Device 287

8.4 The Navigation Problem in Mobile Portals 291
8.4.1 Click-Distance 291
8.4.2 Adaptive Mobile Portals 292
8.4.3 Adaptive Web Navigation 294

8.5 Mobile Search 295
8.5.1 Mobile Search Interfaces 296
8.5.2 Search Engine Support for Mobile Devices 298
8.5.3 Focused Mobile Search 299
8.5.4 Laid Back Mobile Search 300
8.5.5 Mobile Query Log Analysis 301
8.5.6 Personalization of Mobile Search 302
8.5.7 Location-Aware Mobile Search 303

CHAPTER 9  SOCIAL NETWORKS  309

9.1 What is a Social Network? 311
9.1.1 Milgram’s Small-World Experiment 312
9.1.2 Collaboration Graphs 313
9.1.3 Instant Messaging Social Network 314
CONTENTS

9.1.4 The Social Web 314
9.1.5 Social Network Start-Ups 316

9.2 Social Network Analysis 320
9.2.1 Social Network Terminology 320
9.2.2 The Strength of Weak Ties 322
9.2.3 Centrality 322
9.2.4 Web Communities 324
9.2.5 Pajek: Large Network Analysis Software 326

9.3 Peer-to-Peer Networks 326
9.3.1 Centralized P2P Networks 327
9.3.2 Decentralized P2P Networks 328
9.3.3 Hybrid P2P Networks 330
9.3.4 Distributed Hash Tables 331
9.3.5 BitTorrent File Distribution 331
9.3.6 JXTA P2P Search 332
9.3.7 Incentives in P2P Systems 332

9.4 Collaborative Filtering 333
9.4.1 Amazon.com 333
9.4.2 Collaborative Filtering Explained 334
9.4.3 User-Based Collaborative Filtering 335
9.4.4 Item-Based Collaborative Filtering 337
9.4.5 Model-Based Collaborative Filtering 338
9.4.6 Content-Based Recommendation Systems 339
9.4.7 Evaluation of Collaborative Filtering Systems 340
9.4.8 Scalability of Collaborative Filtering Systems 341
9.4.9 A Case Study of Amazon.co.uk 341
9.4.10 The Netflix Prize 342
9.4.11 Some Other Collaborative Filtering Systems 346

9.5 Weblogs (Blogs) 347
9.5.1 Blogrolling 348
9.5.2 Blogspace 348
9.5.3 Blogs for Testing Machine Learning Algorithms 349
9.5.4 Spreading Ideas via Blogs 349
9.5.5 The Real-Time Web and Microblogging 350

9.6 Power-Law Distributions in the Web 352
9.6.1 Detecting Power-Law Distributions 353
9.6.2 Power-Law Distributions in the Internet 355
9.6.3 A Law of Surfing and a Law of Participation 355
9.6.4 The Evolution of the Web via Preferential Attachment 357
9.6.5 The Evolution of the Web as a Multiplicative Process 359
9.6.6 The Evolution of the Web via HOT 360
9.6.7 Small-World Networks 361
9.6.8 The Robustness and Vulnerability of a Scale-Free Network 366

9.7 Searching in Social Networks 369
9.7.1 Social Navigation 369
9.7.2 Social Search Engines 370
9.7.3 Navigation Within Social Networks 373
9.7.4 Navigation Within Small-World Networks 375
9.7.5 Testing Navigation Strategies in Social Networks 379

9.8 Social Tagging and Bookmarking 379