An Introduction to Search Engines and Web Navigation

MARK LEVENE

ADDISON-WESLEY
An imprint of Pearson Education
Harlow, England • London • New York • Boston • San Francisco • Toronto
Sydney • Tokyo • Singapore • Hong Kong • Seoul • Taipei • New Delhi
Cape Town • Madrid • Mexico City • Amsterdam • Munich • Paris • Milan
Contents

Figures xv
Tables xix
Preface xxi
Publisher's Acknowledgements xxiv

1 Introduction 1
1.1 Brief Summary of Chapters 2
1.2 Brief History of Hypertext and the Web 2
1.3 Brief History of Search Engines 6

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 13
2.2 Tabular Data versus Web Data 14
2.3 The Structure of the Web 16
2.3.1 Bow-tie structure of the web 16
2.3.2 Small-world structure of the web 18
2.4 Information Seeking on the Web 19
2.4.1 Direct navigation 19
2.4.2 Navigation within a directory 20
2.4.3 Navigation using a search engine 21
2.4.4 Problems with web information seeking 22
2.5 Informational, Navigational and Transactional Queries 23
2.6 Comparing Web Search to Traditional Information Retrieval 24
2.6.1 Recall and precision 25
2.7 Local Site Search versus Global Web Search 27
2.8 The Difference between Search and Navigation 29
Chapter Summary 30
Exercises 31

3 The Problem of Web Navigation 33
3.1 Getting Lost in Hyperspace and the Navigation Problem 34
3.2 How Can the Machine Assist in User Search and Navigation? 37
3.2.1 The potential use of machine learning algorithms 37
3.2.2 The naive Bayes classifier for categorizing web pages 38
3.3 Trails Should Be First-Class Objects 41
3.4 Enter Markov Chains and Two Interpretations of Their Probabilities 44
  3.4.1 Markov chains and the Markov property 44
  3.4.2 Markov chains and the probabilities of following links 44
  3.4.3 Markov chains and the relevance of links 46
3.5 Conflict between Web Site Owner and Visitor 48
3.6 Conflict between Semantics of Web Site and the Business Model 49
Chapter Summary 51
Exercises 52

4 Searching the Web 53
  4.1 The Mechanics of a Typical Search 54
  4.2 Search Engines as Information Gatekeepers of the Web 58
  4.3 Search Engine Wars: Is the Dust Settling? 60
    4.3.1 Competitor number 1: Google 60
    4.3.2 Competitor number 2: Yahoo™ 61
    4.3.3 Competitor number 3: MSN Search 61
    4.3.4 Other competitors 62
  4.4 Statistics from Studies of Search Engine Query Logs 63
    4.4.1 Search engine query logs 63
    4.4.2 Search engine query syntax 64
    4.4.3 The most popular search keywords 65
  4.5 Architecture of a Search Engine 66
    4.5.1 The search index 67
    4.5.2 The query engine 68
    4.5.3 The search interface 68
  4.6 Crawling the Web 69
    4.6.1 Crawling algorithms 70
    4.6.2 Refreshing web pages 71
    4.6.3 The Robots Exclusion Protocol 72
    4.6.4 Spider traps 72
  4.7 What Does It Take to Deliver a Global Search Service? 73
Chapter Summary 74
Exercises 75

5 How Does a Search Engine Work? 77
  5.1 Content Relevance 78
    5.1.1 Processing web pages 79
    5.1.2 Interpreting the query 81
    5.1.3 Term frequency 81
5.1.4 IDF 84
5.1.5 Computing keyword TF-IDF values 85
5.1.6 Caching queries 86
5.1.7 Phrase matching 86
5.1.8 Synonyms 87
5.1.9 Link text 88
5.1.10 URL analysis 88
5.1.11 Date last updated 88
5.1.12 HTML structure weighting 88
5.1.13 Spell checking 89
5.1.14 Home page detection 90

5.2 Link-Based Metrics 90
5.2.1 Referential and informational links 91
5.2.2 Combining link analysis with content relevance 92
5.2.3 Are links the currency of the web? 92
5.2.4 PageRank explained 93
5.2.5 Online computation of PageRank 97
5.2.6 Hyperlink-Induced Topic Search 97
5.2.7 Counting incoming links 100
5.2.8 Weblogs' influence on PageRank 100
5.2.9 Link spam 101
5.2.10 Citation analysis 101
5.2.11 The wide-ranging interest in PageRank 103

5.3 Popularity-Based Metrics 104
5.3.1 Direct Hit's popularity metric 104
5.3.2 Document space modification 105
5.3.3 Using query log data to improve search 106

5.4 Evaluating Search Engines 107
5.4.1 Search engine awards 107
5.4.2 Evaluation metrics 107
5.4.3 Performance measures 108
5.4.4 Test collections 109
5.4.5 Inferring ranking algorithms 110

Chapter Summary 111
Exercises 113

6 Different Types of Search Engines 115

6.1 Directories and Categorization of Web Content 116
6.2 Paid Inclusion and Paid Placement 117
   6.2.1 Paid inclusion 118
   6.2.2 Paid placement and pay-per-click 118
## 6.3 Metasearch

6.3.1 Fusion algorithms 124  
6.3.2 Operational metasearch engines 125  
6.3.3 Clustering search results 127

## 6.4 Personalization

6.4.1 Personalized results tool 130  
6.4.2 Privacy and scalability 131  
6.4.3 Relevance feedback 132  
6.4.4 Personalized PageRank 133  
6.4.5 Outride’s personalized search 135

## 6.5 Question Answering (Q&A) on the Web

6.5.1 Natural language annotations 136  
6.5.2 Open domain Q&A 139  
6.5.3 Semantic headers 140

## 6.6 Image Search

6.6.1 Text-based image search 142  
6.6.2 Content-based image search 143  
6.6.3 Image search for finding location-based information 147

## 6.7 Special-Purpose Search Engines

Chapter Summary 151  
Exercises 152

## 7 Navigating the Web

7.1 Frustration in Web Browsing and Navigation 156  
7.1.1 HTML and web site design 157  
7.1.2 Hyperlinks and surfing 157  
7.1.3 Web site design and usability 158

7.2 Navigation Tools 158  
7.2.1 The basic browser tools 158  
7.2.2 The back and forward buttons 159  
7.2.3 Search engine toolbars 161  
7.2.4 The bookmarks tool 161  
7.2.5 The history list 164  
7.2.6 Identifying web pages 165  
7.2.7Breadcrumb navigation 166  
7.2.8 Hypertext orientation tools 167  
7.2.9 Hypercard programming environment 168

7.3 Navigational Metrics 169  
7.3.1 The potential gain 170  
7.3.2 Structural analysis of a web site 172  
7.3.3 Measuring the usability of web sites 173
7.4 Web Data Mining
7.4.1 Three perspectives on data mining
7.4.2 Measuring the success of a web site
7.4.3 Web analytics
7.4.4 E-metrics
7.4.5 Weblog file analyzers
7.4.6 Identifying the surfer
7.4.7 Sessionizing
7.4.8 Supplementary analyses
7.4.9 Markov chain model of web site navigation
7.4.10 Applications of web usage mining
7.5 The Best Trail Algorithm
7.5.1 Effective view navigation
7.5.2 Web usage mining for personalization
7.5.3 Developing a trail engine
7.6 Visualization that Aids Navigation
7.6.1 Visualization of navigation patterns
7.6.2 Overview diagrams
7.6.3 Web site maps
7.6.4 Fisheye views
7.6.5 Visualizing trails within a web site
7.6.6 Visual search engines
7.6.7 Mapping cyberspace
7.7 Navigation in Virtual and Physical Spaces
7.7.1 Real-world web usage mining
7.7.2 The museum experience recorder
7.7.3 Navigating in the real world
Chapter Summary
Exercises

8 The Mobile Web
8.1 The Paradigm of Mobile Computing
8.1.1 Wireless Markup Language
8.1.2 The i-mode service
8.2 Mobile Web Services
8.2.1 M-commerce
8.2.2 Delivery of personalized news
8.2.3 Delivery of learning resources
8.3 Mobile Device Interfaces
8.3.1 Mobile web browsers
8.3.2 Information seeking on mobile devices
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.3.3</td>
<td>Text entry on mobile devices</td>
<td>225</td>
</tr>
<tr>
<td>8.3.4</td>
<td>Presenting information on a mobile device</td>
<td>226</td>
</tr>
<tr>
<td>8.4</td>
<td>The Navigation Problem in Mobile Portals</td>
<td>230</td>
</tr>
<tr>
<td>8.4.1</td>
<td>Click-distance</td>
<td>231</td>
</tr>
<tr>
<td>8.4.2</td>
<td>Adaptive mobile portals</td>
<td>231</td>
</tr>
<tr>
<td>8.4.3</td>
<td>Adaptive web navigation</td>
<td>233</td>
</tr>
<tr>
<td>8.5</td>
<td>Mobile Search</td>
<td>234</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Mobile search interfaces</td>
<td>235</td>
</tr>
<tr>
<td>8.5.2</td>
<td>Search engine support for mobile devices</td>
<td>236</td>
</tr>
<tr>
<td>8.5.3</td>
<td>Focused mobile search</td>
<td>238</td>
</tr>
<tr>
<td>8.5.4</td>
<td>Laid-back mobile search</td>
<td>240</td>
</tr>
<tr>
<td>8.5.5</td>
<td>Mobile query log analysis</td>
<td>241</td>
</tr>
<tr>
<td>8.5.6</td>
<td>Personalization of mobile search</td>
<td>241</td>
</tr>
<tr>
<td>8.5.7</td>
<td>Location-aware mobile search</td>
<td>242</td>
</tr>
<tr>
<td>Chapter Summary</td>
<td></td>
<td>243</td>
</tr>
<tr>
<td>Exercises</td>
<td></td>
<td>245</td>
</tr>
</tbody>
</table>

## 9 Social Networks

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>What Is a Social Network?</td>
<td>248</td>
</tr>
<tr>
<td>9.1.1</td>
<td>Milgram's small-world experiment</td>
<td>249</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Collaboration graphs</td>
<td>250</td>
</tr>
<tr>
<td>9.1.3</td>
<td>Instant messaging social network</td>
<td>251</td>
</tr>
<tr>
<td>9.1.4</td>
<td>The social web</td>
<td>251</td>
</tr>
<tr>
<td>9.1.5</td>
<td>Social network startups</td>
<td>253</td>
</tr>
<tr>
<td>9.2</td>
<td>Social Network Analysis</td>
<td>255</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Social network terminology</td>
<td>255</td>
</tr>
<tr>
<td>9.2.2</td>
<td>The strength of weak ties</td>
<td>257</td>
</tr>
<tr>
<td>9.2.3</td>
<td>Centrality</td>
<td>257</td>
</tr>
<tr>
<td>9.2.4</td>
<td>Web communities</td>
<td>259</td>
</tr>
<tr>
<td>9.2.5</td>
<td>Pajek – large-network analysis software</td>
<td>261</td>
</tr>
<tr>
<td>9.3</td>
<td>Peer-to-Peer Networks</td>
<td>261</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Centralized P2P networks</td>
<td>262</td>
</tr>
<tr>
<td>9.3.2</td>
<td>Decentralized P2P networks</td>
<td>262</td>
</tr>
<tr>
<td>9.3.3</td>
<td>Hybrid P2P networks</td>
<td>264</td>
</tr>
<tr>
<td>9.3.4</td>
<td>BitTorrent file distribution</td>
<td>265</td>
</tr>
<tr>
<td>9.3.5</td>
<td>JXTA P2P search</td>
<td>266</td>
</tr>
<tr>
<td>9.4</td>
<td>Collaborative Filtering</td>
<td>266</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Amazon.com</td>
<td>266</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Collaborative filtering explained</td>
<td>267</td>
</tr>
<tr>
<td>9.4.3</td>
<td>User-based collaborative filtering</td>
<td>268</td>
</tr>
</tbody>
</table>