## Contents

**Preface** ix

**Chapter 1. Introduction to Web Search Engines**
- 1.1 A Short History of Information Retrieval 1
- 1.2 An Overview of Traditional Information Retrieval 5
- 1.3 Web Information Retrieval 9

**Chapter 2. Crawling, Indexing, and Query Processing**
- 2.1 Crawling 15
- 2.2 The Content Index 19
- 2.3 Query Processing 21

**Chapter 3. Ranking Webpages by Popularity**
- 3.1 The Scene in 1998 25
- 3.2 Two Theses 26
- 3.3 Query-Independence 30

**Chapter 4. The Mathematics of Google’s PageRank**
- 4.1 The Original Summation Formula for PageRank 32
- 4.2 Matrix Representation of the Summation Equations 33
- 4.3 Problems with the Iterative Process 34
- 4.4 A Little Markov Chain Theory 36
- 4.5 Early Adjustments to the Basic Model 36
- 4.6 Computation of the PageRank Vector 39
- 4.7 Theorem and Proof for Spectrum of the Google Matrix 45

**Chapter 5. Parameters in the PageRank Model**
- 5.1 The $\alpha$ Factor 47
- 5.2 The Hyperlink Matrix $H$ 48
- 5.3 The Teleportation Matrix $E$ 49

**Chapter 6. The Sensitivity of PageRank**
- 6.1 Sensitivity with respect to $\alpha$ 57
6.2 Sensitivity with respect to $H$  
6.3 Sensitivity with respect to $v^T$  
6.4 Other Analyses of Sensitivity  
6.5 Sensitivity Theorems and Proofs  

Chapter 7. The PageRank Problem as a Linear System  
7.1 Properties of $(I - \alpha S)$  
7.2 Properties of $(I - \alpha H)$  
7.3 Proof of the PageRank Sparse Linear System  

Chapter 8. Issues in Large-Scale Implementation of PageRank  
8.1 Storage Issues  
8.2 Convergence Criterion  
8.3 Accuracy  
8.4 Dangling Nodes  
8.5 Back Button Modeling  

Chapter 9. Accelerating the Computation of PageRank  
9.1 An Adaptive Power Method  
9.2 Extrapolation  
9.3 Aggregation  
9.4 Other Numerical Methods  

Chapter 10. Updating the PageRank Vector  
10.1 The Two Updating Problems and their History  
10.2 Restarting the Power Method  
10.3 Approximate Updating Using Approximate Aggregation  
10.4 Exact Aggregation  
10.5 Exact vs. Approximate Aggregation  
10.6 Updating with Iterative Aggregation  
10.7 Determining the Partition  
10.8 Conclusions  

Chapter 11. The HITS Method for Ranking Webpages  
11.1 The HITS Algorithm  
11.2 HITS Implementation  
11.3 HITS Convergence  
11.4 HITS Example  
11.5 Strengths and Weaknesses of HITS  
11.6 HITS's Relationship to Bibliometrics  
11.7 Query-Independent HITS  
11.8 Accelerating HITS  
11.9 HITS Sensitivity
# Chapter 12. Other Link Methods for Ranking Webpages

12.1 SALSA 131
12.2 Hybrid Ranking Methods 135
12.3 Rankings based on Traffic Flow 136

# Chapter 13. The Future of Web Information Retrieval

13.1 Spam 139
13.2 Personalization 142
13.3 Clustering 142
13.4 Intelligent Agents 143
13.5 Trends and Time-Sensitive Search 144
13.6 Privacy and Censorship 146
13.7 Library Classification Schemes 147
13.8 Data Fusion 148

# Chapter 14. Resources for Web Information Retrieval

14.1 Resources for Getting Started 149
14.2 Resources for Serious Study 150

# Chapter 15. The Mathematics Guide

15.1 Linear Algebra 153
15.2 Perron–Frobenius Theory 167
15.3 Markov Chains 175
15.4 Perron Complementation 186
15.5 Stochastic Complementation 192
15.6 Censoring 194
15.7 Aggregation 195
15.8 Disaggregation 198

# Chapter 16. Glossary

201

Bibliography 207

Index 219