

# Understanding Search Engines

## Mathematical Modeling and Text Retrieval

Michael W. Berry

*University of Tennessee  
Knoxville, Tennessee*

Murray Browne

*University of Tennessee  
Knoxville, Tennessee*

**SOFTWARE • ENVIRONMENTS • TOOLS**

**siam**

Society for Industrial and Applied Mathematics  
Philadelphia

# Contents

|                                        |           |
|----------------------------------------|-----------|
| Preface to the Second Edition          | xi        |
| Preface to the First Edition           | xv        |
| <b>1 Introduction</b>                  | <b>1</b>  |
| 1.1 Document File Preparation          | 2         |
| 1.1.1 Manual Indexing                  | 2         |
| 1.1.2 File Cleanup                     | 3         |
| 1.2 Information Extraction             | 4         |
| 1.3 Vector Space Modeling              | 4         |
| 1.4 Matrix Decompositions              | 6         |
| 1.5 Query Representations              | 7         |
| 1.6 Ranking and Relevance Feedback     | 8         |
| 1.7 Searching by Link Structure        | 9         |
| 1.8 User Interface                     | 9         |
| 1.9 Book Format                        | 10        |
| <b>2 Document File Preparation</b>     | <b>11</b> |
| 2.1 Document Purification and Analysis | 12        |
| 2.1.1 Text Formatting                  | 13        |
| 2.1.2 Validation                       | 14        |
| 2.2 Manual Indexing                    | 14        |
| 2.3 Automatic Indexing                 | 16        |
| 2.4 Item Normalization                 | 19        |
| 2.5 Inverted File Structures           | 21        |
| 2.5.1 Document File                    | 22        |
| 2.5.2 Dictionary List                  | 23        |
| 2.5.3 Inversion List                   | 24        |

|          |                                       |           |
|----------|---------------------------------------|-----------|
| 2.5.4    | Other File Structures                 | 26        |
| <b>3</b> | <b>Vector Space Models</b>            | <b>29</b> |
| 3.1      | Construction                          | 29        |
| 3.1.1    | Term-by-Document Matrices             | 30        |
| 3.1.2    | Simple Query Matching                 | 32        |
| 3.2      | Design Issues                         | 34        |
| 3.2.1    | Term Weighting                        | 34        |
| 3.2.2    | Sparse Matrix Storage                 | 38        |
| 3.2.3    | Low-Rank Approximations               | 40        |
| <b>4</b> | <b>Matrix Decompositions</b>          | <b>45</b> |
| 4.1      | QR Factorization                      | 45        |
| 4.2      | Singular Value Decomposition          | 51        |
| 4.2.1    | Low-Rank Approximations               | 55        |
| 4.2.2    | Query Matching                        | 55        |
| 4.2.3    | Software                              | 57        |
| 4.3      | Semidiscrete Decomposition            | 58        |
| 4.4      | Updating Techniques                   | 59        |
| <b>5</b> | <b>Query Management</b>               | <b>63</b> |
| 5.1      | Query Binding                         | 63        |
| 5.2      | Types of Queries                      | 65        |
| 5.2.1    | Boolean Queries                       | 65        |
| 5.2.2    | Natural Language Queries              | 66        |
| 5.2.3    | Thesaurus Queries                     | 66        |
| 5.2.4    | Fuzzy Queries                         | 67        |
| 5.2.5    | Term Searches                         | 67        |
| 5.2.6    | Probabilistic Queries                 | 68        |
| <b>6</b> | <b>Ranking and Relevance Feedback</b> | <b>71</b> |
| 6.1      | Performance Evaluation                | 72        |
| 6.1.1    | Precision                             | 73        |
| 6.1.2    | Recall                                | 73        |
| 6.1.3    | Average Precision                     | 74        |
| 6.1.4    | Genetic Algorithms                    | 75        |
| 6.2      | Relevance Feedback                    | 75        |

|                                                  |            |
|--------------------------------------------------|------------|
| <b>7 Searching by Link Structure</b>             | <b>77</b>  |
| 7.1 HITS Method . . . . .                        | 79         |
| 7.1.1 HITS Implementation . . . . .              | 81         |
| 7.1.2 HITS Summary . . . . .                     | 82         |
| 7.2 PageRank Method . . . . .                    | 84         |
| 7.2.1 PageRank Adjustments . . . . .             | 85         |
| 7.2.2 PageRank Implementation . . . . .          | 87         |
| 7.2.3 PageRank Summary . . . . .                 | 88         |
| <b>8 User Interface Considerations</b>           | <b>89</b>  |
| 8.1 General Guidelines . . . . .                 | 89         |
| 8.2 Search Engine Interfaces . . . . .           | 90         |
| 8.2.1 Form Fill-in . . . . .                     | 91         |
| 8.2.2 Display Considerations . . . . .           | 92         |
| 8.2.3 Progress Indication . . . . .              | 92         |
| 8.2.4 No Penalties for Error . . . . .           | 93         |
| 8.2.5 Results . . . . .                          | 93         |
| 8.2.6 Test and Retest . . . . .                  | 94         |
| 8.2.7 Final Considerations . . . . .             | 95         |
| <b>9 Further Reading</b>                         | <b>97</b>  |
| 9.1 General Textbooks on IR . . . . .            | 97         |
| 9.2 Computational Methods and Software . . . . . | 98         |
| 9.3 Search Engines . . . . .                     | 100        |
| 9.4 User Interfaces . . . . .                    | 100        |
| <b>Bibliography</b>                              | <b>103</b> |
| <b>Index</b>                                     | <b>113</b> |