

# Contents

|          |                          |          |
|----------|--------------------------|----------|
| <b>1</b> | <b>Introduction</b>      | <b>1</b> |
| 1.1      | Motivation               | 1        |
| 1.2      | Definition of Scope      | 2        |
| 1.3      | Organization and Content | 3        |
| 1.3.1    | Chapter Outline          | 3        |
| 1.3.2    | Publications             | 3        |

## Part I Foundations and Use Cases

|          |                                             |          |
|----------|---------------------------------------------|----------|
| <b>2</b> | <b>Research Foundations</b>                 | <b>7</b> |
| 2.1      | Introduction                                | 7        |
| 2.2      | Information Retrieval and Search Technology | 8        |
| 2.2.1    | Vector Space Model Basics                   | 8        |
| 2.2.1.1  | Reducing Term Vector Dimensionality         | 10       |
| 2.2.1.2  | Textual Search Based on VSM                 | 12       |
| 2.2.2    | On Search Engine Anatomy                    | 12       |
| 2.2.2.1  | Inverted Index Organization                 | 13       |
| 2.2.2.2  | Towards Semantic Search Engines             | 14       |
| 2.3      | Data Mining                                 | 16       |
| 2.3.1    | Supervised Learning                         | 17       |
| 2.3.1.1  | Decision Tree Induction                     | 17       |
| 2.3.1.2  | Neural Networks                             | 18       |
| 2.3.1.3  | Naïve-Bayes Classifiers                     | 19       |
| 2.3.1.4  | Support Vector Machines                     | 20       |
| 2.3.2    | Unsupervised Learning                       | 20       |
| 2.3.2.1  | $k$ -Means Clustering                       | 21       |
| 2.3.2.2  | Expectation Maximization                    | 22       |
| 2.3.2.3  | Self-Organizing Maps                        | 23       |
| 2.3.2.4  | Density-Based Clustering                    | 24       |
| 2.4      | Natural Language Processing and Text Mining | 26       |
| 2.4.1    | Text-Based Classification and Clustering    | 27       |
| 2.4.2    | Named Entity Recognition                    | 27       |

|         |                                                                    |    |
|---------|--------------------------------------------------------------------|----|
| 2.4.3   | Keyword Extraction . . . . .                                       | 28 |
| 2.4.4   | Identification of Word Collocations . . . . .                      | 29 |
| 2.5     | Social Network Analysis . . . . .                                  | 29 |
| 2.5.1   | SNA in the Context of Competitive Strategic Intelligence . . . . . | 30 |
| 2.5.2   | Network Centrality Measures . . . . .                              | 31 |
| 2.5.2.1 | Degree Centrality . . . . .                                        | 32 |
| 2.5.2.2 | Closeness Centrality . . . . .                                     | 32 |
| 2.5.2.3 | Betweenness Centrality . . . . .                                   | 33 |
| 2.5.2.4 | Eigenvector Centrality . . . . .                                   | 34 |
| 2.5.3   | Statistical Properties of Social Networks . . . . .                | 36 |
| 2.5.3.1 | On Random Graphs . . . . .                                         | 37 |
| 2.5.3.2 | Small-World Models . . . . .                                       | 37 |
| 2.5.3.3 | Preferential Attachment Models . . . . .                           | 38 |
| 2.5.4   | Conclusion . . . . .                                               | 39 |
| 2.6     | Sentiment Detection . . . . .                                      | 40 |
| 2.6.1   | Issues and Challenges . . . . .                                    | 41 |
| 2.6.1.1 | Impact of Context on Polarity . . . . .                            | 41 |
| 2.6.1.2 | Topic Identification . . . . .                                     | 42 |
| 2.6.1.3 | Language, Grammar, and Idiomatic Expressions . . . . .             | 42 |
| 2.6.2   | Sentiment Detection Techniques . . . . .                           | 44 |
| 2.6.2.1 | Machine Learning and Lexical Approaches . . . . .                  | 44 |
| 2.6.2.2 | Enhancing Sentiment Detection through NLP Approaches . . . . .     | 46 |
| 2.6.3   | Going beyond Research . . . . .                                    | 48 |
| 2.6.4   | Conclusion and Outlook . . . . .                                   | 48 |
| 3       | <b>Competitive Intelligence Capturing Systems . . . . .</b>        | 51 |
| 3.1     | Introduction . . . . .                                             | 51 |
| 3.2     | Common Platform Architecture . . . . .                             | 54 |
| 3.2.1   | Document Collection Layer . . . . .                                | 54 |
| 3.2.2   | Analysis and Semantic Annotation Layer . . . . .                   | 54 |
| 3.2.3   | Reporting and User Interface . . . . .                             | 55 |
| 3.3     | Typical Analysis Tasks . . . . .                                   | 55 |
| 3.4     | Survey of Existing Systems . . . . .                               | 58 |
| 4       | <b>Research Topics and Applications . . . . .</b>                  | 63 |
| 4.1     | Introduction . . . . .                                             | 63 |
| 4.2     | Trend and Topic Analysis . . . . .                                 | 63 |
| 4.2.1   | Reputation Monitoring Platform . . . . .                           | 64 |
| 4.2.1.1 | Analysis Services . . . . .                                        | 65 |
| 4.2.1.2 | Research Contributions . . . . .                                   | 69 |
| 4.2.2   | Exploratory Customer Feedback Analysis . . . . .                   | 69 |
| 4.2.2.1 | Platform Components . . . . .                                      | 70 |

|                                      |                                                                                                  |     |
|--------------------------------------|--------------------------------------------------------------------------------------------------|-----|
| 4.2.2.2                              | Conclusion .....                                                                                 | 72  |
| 4.2.3                                | Temporal Topic Evolution .....                                                                   | 72  |
| 4.2.4                                | Analysis of Social Tagging Behavior .....                                                        | 73  |
| 4.3                                  | Automated News Extraction .....                                                                  | 74  |
| 4.3.1                                | Particle Swarm Optimization for Learning .....                                                   | 78  |
| 4.3.1.1                              | PSO Training Procedure .....                                                                     | 78  |
| 4.3.1.2                              | Evaluation and Application .....                                                                 | 80  |
| 4.3.2                                | Learning Based on Classification Models .....                                                    | 80  |
| 4.3.2.1                              | Block Labelling .....                                                                            | 81  |
| 4.3.2.2                              | Classifier Training and Evaluation .....                                                         | 82  |
| 4.3.2.3                              | Feature Impact Assessment .....                                                                  | 83  |
| 4.4                                  | Discovery of Technology Synergies .....                                                          | 84  |
| 4.4.1                                | Semantic Similarity between Named Entities .....                                                 | 85  |
| 4.4.1.1                              | Use of Background Knowledge .....                                                                | 85  |
| 4.4.1.2                              | Creating Semantic Fingerprints .....                                                             | 86  |
| 4.4.1.3                              | Computing Fingerprint Similarity .....                                                           | 87  |
| 4.4.1.4                              | Training and Evaluation .....                                                                    | 87  |
| 4.4.2                                | Web 2.0 Leverage in Finding Technology Synergies .....                                           | 88  |
| 4.4.2.1                              | Characteristics of Synergetic Technologies .....                                                 | 89  |
| 4.4.2.2                              | Building Classifiers .....                                                                       | 90  |
| 4.4.2.3                              | Empirical Evaluation .....                                                                       | 91  |
| 4.5                                  | Concluding Remarks .....                                                                         | 93  |
| 5                                    | <b>Conclusion</b> .....                                                                          | 95  |
| 5.1                                  | Summary .....                                                                                    | 95  |
| 5.2                                  | Outlook and Future Directions .....                                                              | 96  |
|                                      | <b>References</b> .....                                                                          | 99  |
| <b>Part II Selected Publications</b> |                                                                                                  |     |
| 6                                    | <b>Research Papers – Section 4.2</b> .....                                                       | 109 |
|                                      | Towards Automated Reputation and Brand Monitoring on the Web .....                               | 111 |
|                                      | Mining and Exploring Customer Feedback Using Language Models<br>and Treemaps .....               | 121 |
| 7                                    | <b>Research Papers – Section 4.3</b> .....                                                       | 135 |
|                                      | Content Extraction from News Pages Using Particle Swarm<br>Optimization .....                    | 137 |
|                                      | Distilling Informative Content from HTML News Pages Using<br>Machine Learning Classifiers .....  | 151 |
| 8                                    | <b>Research Papers – Section 4.4</b> .....                                                       | 167 |
|                                      | Automatic Computation of Semantic Proximity Using Taxonomic<br>Knowledge .....                   | 169 |
|                                      | Leveraging Sources of Collective Wisdom on the Web for<br>Discovering Technology Synergies ..... | 189 |