Graph theory and the fields of natural language processing and information retrieval are well-studied disciplines. Traditionally, these areas have been perceived as distinct, with different algorithms, different applications, and different potential end-users. However, recent research has shown that these disciplines are intimately connected, with much variety in the way that natural language processing and information retrieval applications find efficient solutions within graph-theoretical frameworks.

This book is a comprehensive description of the use of graph-based algorithms for natural language processing and information retrieval. It brings together topics as diverse as lexical semantics, text summarization, text mining, ontology construction, text classification, and text retrieval, which are connected by the common underlying theme of the use of graph-theoretical methods for text- and information-processing tasks. Readers will gain a firm understanding of the major methods and applications in natural language processing and information retrieval that rely on graph-based representations and algorithms.

Rada Mihalcea is an Associate Professor in the Department of Computer Science and Engineering at the University of North Texas, where she leads the Language and Information Technologies research group. In 2009, she received the Presidential Early Career Award for Scientists and Engineers, awarded by President Barack Obama. She served on the editorial board of several journals, including *Computational Linguistics, Journal of Natural Language Engineering, and Language Resources and Evaluations*, and she cochaired the Empirical Methods in Natural Language Processing Conference in 2009 and the Association for Computational Linguistics Conference in 2011. She has been published in *IEEE Intelligent Systems, Journal of Natural Language Engineering, Journal of Machine Translation, Computational Intelligence, International Journal of Semantic Computing*, and *Artificial Intelligence Magazine*.

Dragomir Radev is a Professor in the School of Information, the Department of Electrical Engineering and Computer Science, and the Department of Linguistics at the University of Michigan, where he is the leader of the Computational Linguistics and Information Retrieval (CLAIR) research group. He has had more than 100 publications in conferences and journals such as *Communications of the Association for Computing Machinery (ACM), Journal of Artificial Intelligence Research, Bioinformatics, Computational Linguistics, Information Processing and Management*, and *American Journal of Political Science*, among others. He is on the editorial boards of *Information Retrieval, Journal of Natural Language Engineering, and Journal of Artificial Intelligence Research*. Professor Radev is an ACM distinguished scientist as well as coach of the U.S. high school team in computational linguistics. He is also an Adjunct Professor in computer science at Columbia University.
GRAPH-BASED NATURAL LANGUAGE PROCESSING AND INFORMATION RETRIEVAL

RADA MIHALCEA
University of North Texas,
Department of Computer Science and Engineering

DRAGOMIR RADEV
University of Michigan,
School of Information
Department of Electrical Engineering and Computer Science
Department of Linguistics
## Contents

Introduction

0.1 Background 3
0.2 Book Organization 4
0.3 Acknowledgments 7

Part I. Introduction to Graph Theory

1 Notations, Properties, and Representations 11
  1.1 Graph Terminology and Notations 11
  1.2 Graph Properties 13
  1.3 Graph Types 14
  1.4 Representing Graphs as Matrices 15
  1.5 Using Matrices to Compute Graph Properties 16
  1.6 Representing Graphs as Linked Lists 17
  1.7 Eigenvalues and Eigenvectors 18

2 Graph-Based Algorithms 20
  2.1 Depth-First Graph Traversal 20
  2.2 Breadth-First Graph Traversal 22
  2.3 Minimum Spanning Trees 23
  2.4 Shortest-Path Algorithms 26
  2.5 Cuts and Flows 29
  2.6 Graph Matching 31
  2.7 Dimensionality Reduction 32
  2.8 Stochastic Processes on Graphs 34
  2.9 Harmonic Functions 38
  2.10 Random Walks 40
  2.11 Spreading Activation 41
  2.12 Electrical Interpretation of Random Walks 42
  2.13 Power Method 44
2.14 Linear Algebra Methods for Computing Harmonic Functions
2.15 Method of Relaxations
2.16 Monte Carlo Methods

Part II. Networks
3 Random Networks
3.1 Networks and Graphs
3.2 Random Graphs
3.3 Degree Distributions
3.4 Power Laws
3.5 Zipf’s Law
3.6 Preferential Attachment
3.7 Giant Component
3.8 Clustering Coefficient
3.9 Small Worlds
3.10 Assortativity
3.11 Centrality
3.12 Degree Centrality
3.13 Closeness Centrality
3.14 Betweenness Centrality
3.15 Network Example
3.16 Dynamic Processes: Percolation
3.17 Strong and Weak Ties
3.18 Assortative Mixing
3.19 Structural Holes

4 Language Networks
4.1 Co-Occurrence Networks
4.2 Syntactic Dependency Networks
4.3 Semantic Networks
4.4 Similarity Networks

Part III. Graph-Based Information Retrieval
5 Link Analysis for the World Wide Web
5.1 The Web as a Graph
5.2 PageRank
5.3 Undirected Graphs
5.4 Weighted Graphs
## Contents

5.5 Combining PageRank with Content Analysis | 97
5.6 Topic-Sensitive Link Analysis | 97
5.7 Query-Dependent Link Analysis | 100
5.8 Hyperlinked-Induced Topic Search | 101
5.9 Document Reranking with Induced Links | 103

6 Text Clustering | 106
6.1 Graph-Based Clustering | 108
6.2 Spectral Methods | 111
6.3 The Fiedler Method | 113
6.4 The Kernighan–Lin Method | 114
6.5 Betweenness-Based Clustering | 115
6.6 Min-Cut Clustering | 117
6.7 Text Clustering Using Random Walks | 119

Part IV. Graph-Based Natural Language Processing

7 Semantics | 123
7.1 Semantic Classes | 123
7.2 Synonym Detection | 125
7.3 Semantic Distance | 126
7.4 Textual Entailment | 129
7.5 Word-Sense Disambiguation | 131
7.6 Name Disambiguation | 134
7.7 Sentiment and Subjectivity | 135

8 Syntax | 140
8.1 Part-of-Speech Tagging | 140
8.2 Dependency Parsing | 141
8.3 Prepositional-Phrase Attachment | 144
8.4 Co-Reference Resolution | 146

9 Applications | 149
9.1 Summarization | 149
9.2 Semi-supervised Passage Retrieval | 150
9.3 Keyword Extraction | 154
9.4 Topic Identification | 156
9.5 Topic Segmentation | 161
9.6 Discourse | 162
9.7 Machine Translation | 165
Contents

9.8 Cross-Language Information Retrieval 166
9.9 Information Extraction 169
9.10 Question Answering 171
9.11 Term Weighting 174

Bibliography 179
Index 191