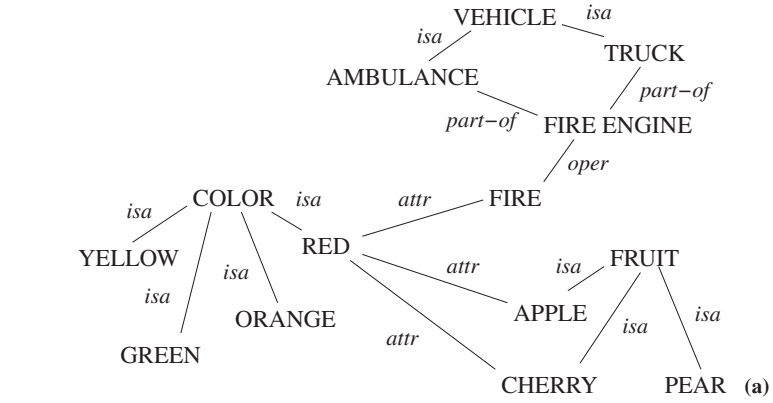
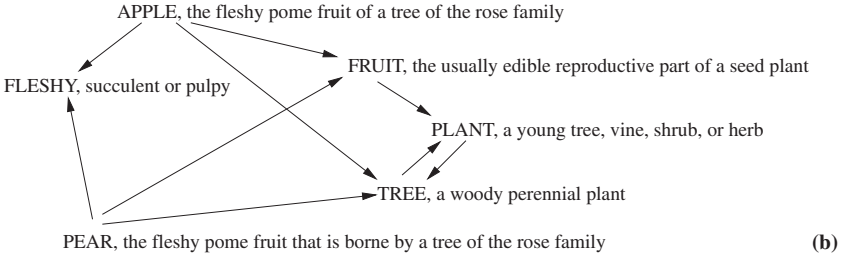
Introduction

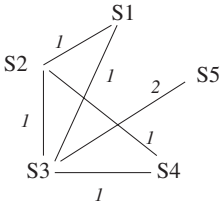
Graph theory is a well-studied discipline as are the fields of natural language processing and information retrieval. Traditionally, these areas of study have been perceived as distinct, with different algorithms, different applications, and different potential end-users. However, as recent research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

In a cohesive text, language units – whether they are words, phrases, or entire sentences – are connected through various relationships, which contribute to the overall meaning and maintain the cohesive structure and discourse unity of the text. Since the early stages of artificial intelligence, associative or semantic networks have been proposed as representations that enable the storage of such language units and their interconnecting relationships, which allow for a variety of inference and reasoning processes that simulate functionalities of the human mind (Sowa 1983). The symbolic structures that emerge from these representations correspond naturally to graphs – in which text constituents are represented as vertices and their interconnecting relationships form the edges in the graph.

Many text-processing applications can be modeled by means of a graph. These data structures have the capability to encode naturally the meaning and structure of a cohesive text and to follow closely the associative or semantic memory representations. For instance, Figure 0.1 is an example of graph representations of language units and their interconnecting relationships: 0.1(a) (adapted from Collins and Loftus [1975] shows a network of concepts related by semantic relations – simulating a fragment of human memory on which reasoning and inferences about various concepts represented in the network can be performed; 0.1(b) shows a network with a similar structure, this time automatically derived via definitional links in a dictionary; and 0.1(c) represents a graph of the cohesive structure of a text by encoding similarity relationships among textual units.







S1: Apples are mentioned throughout most of recorded human history.

S2: The generic name Malus was derived from the word *malus*, or bad, referring to Eve picking an apple.

S3: However, some scholars think the fig, and not the apple, was the forbidden fruit.

S4: The fig is also the first tree mentioned in the story of Adam and Eve.

S5: There are some scholars who think the apricot is a more likely candidate because it was an abundant fruit. **(c)**

Figure 0.1. Graph representations of textual units and their interconnecting relationships.

This book is a comprehensive description of the use of graph-based algorithms for natural language processing and information retrieval. It brings together areas as diverse as lexical semantics, text summarization, text mining, ontology construction, text classification, and information retrieval, which are connected by the common underlying theme of using graph-theoretical methods for text- and information-processing tasks.

The book complements existing textbooks in natural language processing and information retrieval such as Manning and Schütze (1999); Jurafsky and Martin (2009); and Manning, Raghavan, and Schütze (2008); recent related books also cover the topic of network theory, including Kleinberg and Easley (2010), Jackson (2009), and Newman (2010).

Readers of this book 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. Our objective is that readers will have sufficient understanding to make informed decisions about using graph-based algorithms in text-processing applications in the future and to recognize opportunities for advancing the state of the art in natural language processing and information retrieval problems by the application of graph-theoretical algorithms.

0.1. Background

The idea of networks (or graphs[1]) as mental representations for language units and the relationships among them originates with the early work in psychology (Freud 1901) and psycholinguistics (Schvaneveldt 1989; Spitzer 1999). Theories of semantic or associative memory (Quillian 1968) and connectionism (Fodor and Pylyshyn 1988) initially emerged in cognitive psychology as models for human language representation and reasoning; since then, they have been applied to various computer-based applications.

The semantic memory developed by Quillian (1968) was proposed as a representational format that enables the storage of the *meaning* of words or other cognitive units and then makes the use of such meaning possible for other applications. The memory model proposed by Quillian consists of a mass of nodes connected by associative links, which can be either simple associations or more complex attributes of the cognitive units. The model was introduced as a representation of "the memory that a person calls upon in his everyday language behavior" (Quillian 1968), which allows for associations and configurations as varied and rich as ideas expressed by humans in natural language.

Using his semantic-memory model, Quillian demonstrated how a properly constructed semantic network that encodes word types and tokens, as well as relationships among them derived via dictionary definitions, can

[1] Except when noted otherwise, throughout this book, the notions of *graph* and *network* are used interchangeably.

be used successfully to identify the meaning of words, to find nontrivial semantic paths between words and ideas, and even to generate natural language texts starting with a network of concepts and relationships. His early work in semantic-memory models and computer-feasible representations established the foundation for more general theories of text understanding.

Starting with the seminal work of Quillian (1968) on theories of semantic memory, associative or semantic networks have been used as the underlying model for several applications in language processing, including word-sense disambiguation, automated reasoning, text generation, information retrieval, and text summarization.

A major impediment faced by early researchers in this area was the complexity of the resulting structures, which was not always supported by the underlying computer hardware. Consequently, limitations on the size of applications attempted with these approaches resulted. For similar reasons, the evaluation of such algorithms was mostly performed on toy-size problems (e.g., Quillian [1968] evaluated his proposed word-sense disambiguation algorithm on nineteen ambiguous words), and the scalability of the models was rarely evaluated, if ever.

The situation has changed in recent years, however, and the area of graph-based representations and algorithms applied to natural language processing and information retrieval has experienced tremendous growth. Graph-ranking algorithms (e.g., HITS [Kleinberg 1999] and PageRank [Brin and Page 1998]) have been used successfully on a large scale for analysis of the link structure of the World Wide Web (Brin and Page 1998). A series of successful workshops that began in 2006, centered on the theme of graph-based natural language processing, reached its fifth edition, as follows: Mihalcea and Radev 2006, New York; Biemann et al. 2007, Rochester; Matveeva et al. 2008, Manchester; Chowdhury et al. 2009, Singapore; and Banea et al. 2010, Uppsala. A growing number of publications on the use of graphs for text processing (i.e., more than four thousand papers that address the concept of *graph* in the Association for Computational Linguistics Anthology[2]) is also indicative of the increased interest in this area.

0.2. Book Organization

The material covered in this book is divided into four parts: the introductory part explains relevant graph representations and algorithms; the second part

[2] Available at http://acl.ldc.upenn.edu.

discusses networks, including random networks and language networks; the third part presents the application of graphs to information retrieval; and the fourth part addresses the use of graphs in natural language processing.[3]

0.2.1. Part I: Introduction to Graph Theory

The first part of the book is an overview of graph-based representations and algorithms, focusing on those methods that have been successful in the fields of natural language processing and information retrieval.

Notations, Properties, and Representations

Chapter 1 introduces the most commonly used graph notations, which are used as a reference throughout the book. The chapter also discusses the most important properties of graphs: diameter; clustering coefficient; degree distribution and average shortest path; and graph types and classifications, including random graphs, regular graphs, small-world graphs, and scale-free graphs. Finally, Chapter 1 describes representations that can be used in practical implementations, including adjacency lists and matrices.

Graph-Based Algorithms

Chapter 2 discusses the most frequently used graph-theoretical algorithms, emphasizing those that are relevant to the text and information processing applications addressed in Parts III and IV. Included in the chapter are graph-traversal, including depth-first and breadth-first strategies, minimum path length, and minimum spanning trees; flow-on graphs, including algorithms for min-cut/max-flow; graph matching; random walks with harmonic functions and electrical networks; and linear algebra on graphs.

0.2.2. Part II: Networks

Part II addresses random networks and naturally occurring networks, focusing specifically on networks that occur in linguistic structures.

Random Networks

Chapter 3 discusses random networks and naturally occurring networks. Their properties as well as algorithms for computing them are described.

[3] The book does not include Hidden Markov Models (HMM), Conditional Random Fields (CRF), graphical models, trees in general, or neural networks.

Language Networks

Chapter 4 discusses language networks from an empirical perspective. It covers co-occurrence networks, syntactic and semantic networks, similarity networks, and networks of summaries and machine translation. Language networks are used in a number of methods and applications in natural language processing and information retrieval, such as syntax, semantics, and summarization.

0.2.3. Part III: Graph-Based Information Retrieval

Part III addresses the application of graph-based algorithms to problems relevant to the field of information retrieval, focusing on search engines, document reranking, and text classification and clustering.

Link Analysis for the World Wide Web

Chapter 5 addresses link-analysis methods used by search engines (e.g., PageRank and HITS), and discusses topics relevant to their application, including method stability, the combination of link-based and content-based models, topic-sensitive ranking, and query-dependent link analysis.

Text Clustering

Chapter 6 describes text clustering using a variety of graph-based algorithms, including the Fiedler method, Kernighan–Lin method, text categorization with min-cuts, betweenness, and random walks.

0.2.4. Part IV: Graph-Based Natural Language Processing

Part IV covers the application of graph-based methods to natural language processing, addressing separately the areas of semantics, syntax, and language-processing applications.

Semantics

Chapter 7 addresses the application of graph-based algorithms to problems in the area of semantics. It covers synonym detection and automatic construction of semantic classes using measures of graph connectivity on graphs built from either raw text or user-contributed resources; measures of

semantic distance on semantic networks, including simple path-length algorithms as well as more complex random-walk methods; textual entailment using graph-matching algorithms on syntactic or semantic graphs; word-sense disambiguation and name disambiguation, including random-walk algorithms and other structural approaches for knowledge-based word-sense disambiguation as well as semi-supervised methods using label propagation on graphs; and sentiment classification using either semi-supervised graph-based learning or prior subjectivity detection with min-cut/max-flow algorithms.

Syntax

Chapter 8 covers the use of graph-theoretical algorithms for tasks in syntax, including part-of-speech tagging using graphs that encode word and tag dependencies; dependency parsing using minimum spanning trees; prepositional attachment using word-dependency distributions induced from random-walk models; and co-reference resolution using graph clustering and min-cut algorithms.

Applications

Chapter 9 is concerned with graph-theoretical methods for text-processing applications. The chapter includes text summarization using graph-centrality methods; keyword extraction and topic identification using random-walk language models; text segmentation using normalized-cut criteria for graph partitioning; graph structure to encode discourse relationships; word graphs for decoding in machine translation and speech processing; random-walk algorithms for translation selection in cross-language information retrieval; and graph representations and patterns on graphs for information extraction and question answering.

0.3. Acknowledgments

We want to thank attendees of the tutorials covering material from this book that we presented at the conferences on Recent Advances in Natural Language Processing (RANLP 2005), Human Language Technologies–North American Chapter of the Association for Computational Linguistics (HLT–NAACL 2006), and Workshop on Spoken Language Technology (SLT 2006). The feedback received during these tutorials helped us give shape to the book in its current version. We also want to thank the students

of the course on Advanced NLP and IR that we taught in Winter 2006 at the University of Michigan, as well as the students in the advanced course on Graph-Based Natural Language Processing taught in Spring 2008 at the University of North Texas. Some of the material in this book was included in the courses on Network Theory at Columbia University, Graph-Based Natural Language Processing at the University of North Texas, Search Engine Technology at Columbia University, and Information Retrieval at the University of Michigan.

We also thank Ahmed Hassan, Al Aho, Amjad abu Jbara, Chen Huang, Chris Manning, Güneş Erkan, Lada Adamic, Mark Newman, Pradeep Muthukrishnan, Prem Ganeshkumar, Rich Caneba, Ryan McDonald, Sara Stolbach, Sasha Blair-Goldensohn, and Vahed Qazvinian for their comments and assistance with this book.

Finally, we express our appreciation to the institutions that partly funded the work of the authors: National Science Foundation grants IIS 0534323, "Collaborative Research: BlogoCenter – Infrastructure for Collecting, Mining and Accessing Blogs"; IIS 0329043, "Probabilistic and Link-Based Methods for Exploiting Very Large Textual Repositories"; and BCS 0527513, "DHB: The Dynamics of Political Representation and Political Rhetoric"; and several National Institutes for Health grants to Dragomir Radev, R01 LM008106, "Representing and Acquiring Knowledge of Genome Regulation" and U54 DA021519 "National Center for Integrative Bioinformatics." Research Grant 003594, "Graph-Based Natural Language Processing" from the Texas Advanced Research Program; a Google grant on "Finding Important Information in Unstructured Text"; and National Science Foundation Grants IIS 0747340, "CAREER: Semantic Interpretation with Monolingual and Crosslingual Evidence," IIS 0917170 "Multilingual Subjectivity and Word Senses," and IIS 1018613, "Building a Large Multilingual Network for Text Processing Applications" were awarded to Rada Mihalcea. Any opinions, findings, and conclusions or recommendations expressed in this book are those of the authors and do not necessarily reflect the views of the National Science Foundation or the other sponsors.

Part I

Introduction to Graph Theory

1

Notations, Properties, and Representations

THIS CHAPTER introduces the most commonly used *graph notations* that are referenced throughout the book. The chapter also covers the most important *properties of graphs*: *diameter*; *clustering coefficient*; *degree distribution* and *average shortest path*; and *graph types and classifications*, including *random graphs*, *regular graphs*, *small-world graphs*, and *scale-free graphs*. Finally, the chapter describes the *representations* that can be used in practical implementations, including adjacency lists and matrices.

1.1. Graph Terminology and Notations

A *graph* is a data structure consisting of a set of vertices connected by a set of edges that can be used to model relationships among the objects in a collection. Graphs are typically studied in the field of *graph theory*, which is an area of study in mathematical research.

GRAPH

GRAPH THE

Figure 1.1 shows the seven bridges in Königsberg that led to the invention of graph theory. The problem is to cross all bridges without crossing any bridge twice. Euler showed in 1741 that it was impossible to do so (Euler 1741).

Formally, a *graph* is defined as a set $G = (V, E)$, where V is a collection of vertices $V = \{V_i, i = 1, n\}$ and E is a collection of edges over V, $E_{ij} = \{(V_i, V_j), V_i \in V, V_j \in V\}$. In alternative terminology, graphs also are referred to as *networks*, vertices as *nodes*, and edges as *links*.

Graphs can be either *directed* or *undirected*, depending on whether a direction of travel is defined over the edges. In a directed graph (or *digraph*), an edge E_{ij} can be traversed from V_i to V_j but not in the other direction; V_i is called the *tail* of the edge and V_j is called the *head*. In an undirected graph, edges can be traversed in both directions. Figure 1.2(a) is an example of an undirected-graph structure consisting of five nodes connected by eight edges. Figure 1.2(b) is a similar graph but with directed edges. In the figures,