Book Reviews

Graph-Based Natural Language Processing and Information Retrieval

Rada Mihalcea and Dragomir Radev
(University of North Texas and University of Michigan)

Cambridge, UK: Cambridge University Press, 2011, viii+192 pp; hardbound,
ISBN 978-0-521-89613-9, $65.00

Reviewed by
Chris Biemann
Technische Universität Darmstadt

Graphs are ubiquitous. There is hardly any domain in which objects and their relations cannot be intuitively represented as nodes and edges in a graph. Graph theory is a well-studied sub-discipline of mathematics, with a large body of results and a large number of efficient algorithms that operate on graphs. Like many other disciplines, the fields of natural language processing (NLP) and information retrieval (IR) also deal with data that can be represented as a graph. In this light, it is somewhat surprising that only in recent years the applicability of graph-theoretical frameworks to language technology became apparent and increasingly found its way into publications in the field of computational linguistics. Using algorithms that take the overall graph structure of a problem into account, rather than characteristics of single objects or (unstructured) sets of objects, graph-based methods have been shown to improve a wide range of NLP tasks. In a short but comprehensive overview of the field of graph-based methods for NLP and IR, Rada Mihalcea and Dragomir Radev list an extensive number of techniques and examples from a wide range of research papers by a large number of authors. This book provides an excellent review of this research area, and serves both as an introduction and as a survey of current graph-based techniques in NLP and IR. Because the few existing surveys in this field concentrate on particular aspects, such as graph clustering (Lancichinetti and Fortunato 2009) or IR (Liu 2006), a textbook on the topic was very much needed and this book surely fills this gap.

The book is organized in four parts and contains a total of nine chapters. The first part gives an introduction to notions of graph theory, and the second part covers natural and random networks. The third part is devoted to graph-based IR, and part IV covers graph-based NLP. Chapter 1 lays the groundwork for the remainder of the book by introducing all necessary concepts in graph theory, including the notation, graph properties, and graph representations. In the second chapter, a glimpse is offered into the plethora of graph-based algorithms that have been developed independently of applications in NLP and IR. Sacrificing depth for breadth, this chapter does a great job in touching on a wide variety of methods, including minimum spanning trees, shortest-path algorithms, cuts and flows, subgraph matching, dimensionality reduction, random walks, spreading activation, and more. Algorithms are explained concisely, using examples, pseudo-code, and/or illustrations, some of which are very well suited for classroom examples. Network theory is presented in Chapter 3. The term *network* is here used to refer to naturally occurring relations, as opposed to *graphs* being generated by an automated process. After presenting the classical Erdős-Rényi random graph model and showing its inadequacy to model power-law degree distributions following Zipf's law, scale-free small-world networks are introduced. Further,

several centrality measures, as well as other topics in network theory, are defined and exemplified.

Establishing the connection to NLP, Chapter 4 introduces networks constructed from natural language. Co-occurrence networks and syntactic dependency networks are examined quantitatively. Results on the structure of semantic networks such as WordNet are presented, as well as a range of similarity networks between lexical units. This chapter will surely inspire the reader to watch out for networks in his/her own data. Chapter 5 turns to link analysis for the Web. The PageRank algorithm is described at length, variants for undirected and weighted graphs are introduced, and the algorithm's application to topic-sensitive analysis and query-dependent link analysis is discussed. This chapter is the only one that touches on core IR, and this is also the only chapter with content that can be found in other textbooks (e.g., Liu 2011). Still, this chapter is an important prerequisite for the chapter on applications. It would have been possible to move the description of the algorithms to Chapter 2, however, omitting this part.

The topic of Chapter 6 is text clustering with graph-based methods, outlining the Fiedler method, the Kernighan–Lin method, min-cut clustering, betweenness, and random walk clustering. After defining measures on cluster quality for graphs, spectral and non-spectral graph clustering methods are briefly introduced. Most of the chapter is to be understood as a presentation of general graph clustering methods rather than their application to language. For this, some representative methods for different core ideas were selected. Part IV on graph-based NLP contains the chapters probably most interesting to readers working in computational linguistics. In Chapter 7, graph-based methods for lexical semantics are presented, including detection of semantic classes, synonym detection using random walks on semantic networks, semantic distance on WordNet, and textual entailment using graph matching. Methods for word sense and name disambiguation with graph clustering and random walks are described. The chapter closes with graph-based methods for sentiment lexicon construction and subjectivity classification.

Graph-based methods for syntactic processing are presented in Chapter 8: an unsupervised part-of-speech tagging algorithm based on graph clustering, minimum spanning trees for dependency parsing, PP-attachment with random walks over syntactic co-occurrence graphs, and coreference resolution with graph cuts. In the final chapter, many of the algorithms introduced in the previous chapters are applied to NLP applications as diverse as summarization, passage retrieval, keyword extraction, topic identification and segmentation, discourse, machine translation, cross-language IR, term weighting, and question answering.

As someone with a background in graph-based NLP, I enjoyed reading this book. The writing style is concise and clear, and the authors succeed in conveying the most important points from an incredibly large number of works, viewed from the graph-based perspective. I also liked the extensive use of examples—throughout, almost half of the space is used for figures and tables illustrating the methods, which some readers might perceive as unbalanced, however. With just under 200 pages and a topic as broad as this, it necessarily follows that many of the presented methods are exemplified and touched upon rather than discussed in great detail. Although this sometimes leads to the situation that some passages can only be understood with background knowledge, it is noteworthy that every chapter includes a section on further reading. In this way, the book serves as an entry point to a deeper engagement with graph-based methods for NLP and IR, and it encourages readers to see their NLP problem from a graph-based view.

For a future edition, however, I have a few wishes: It would be nice if the figures and examples were less detached from the text and explained more thoroughly. At times, it would be helpful to present deeper insights and to connect the methodologies, rather than just presenting them next to each other. Also, some of the definitions in Chapter 2 could be less confusing and structured better.

Because this book emphasizes graph-based aspects for language processing rather than aiming at exhaustively treating the numerous tasks that benefit from graph-based methods, it cannot replace a general introduction to NLP or IR: For students without prior knowledge in NLP and IR, a more guided and focused approach to the topic would be required. The target audience is, rather, NLP researchers and professionals who want to add the graph-based view to their arsenal of methods, and to become inspired by this rapidly growing research area. It is equally suited for people working in graph algorithms to learn about graphs in language as a field of application for their work. I will surely consult this volume in the future to supplement the preparation of lectures because of its comprehensive references and its richness in examples.

References

Lancichinetti, Andrea and Santo Fortunato. 2009. Community detection algorithms: A comparative analysis. *Physical Review E*, 80:056117.

Liu, Bing. 2011. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data* (second edition). Berlin, Springer.

Liu, Yi. 2006. Graph-based learning models for information retrieval: A survey. Available at: `www.cse.msu.edu/∼rongjin/semisupervised/graph.pdf`.

Chris Biemann is Juniorprofessor (assistant professor) for Language Technology at Darmstadt University of Technology. His current research interests include statistical semantics, graph-based methods for unsupervised acquisition, and topic modeling. Biemann's address is UKP lab, Computer Science Department, Hochschulstr. 10, 64289 Darmstadt, Germany; e-mail: `biemann@tk.informatik.tu-darmstadt.de`.