Managing and Mining Graph Data

ADVANCES IN DATABASE SYSTEMS
Volume 40

Series Editors

Ahmed K. Elmagarmid
Purdue University
West Lafayette, IN 47907

Amit P. Sheth
Wright State University
Dayton, OH 45435

For other titles published in this series, please visit www.springer.com/series/5573

Managing and Mining Graph Data

by

Charu C. Aggarwal
IBM T.J. Watson Research Center
Hawthorne, NY, USA

Haixun Wang
Microsoft Research Asia
Beijing, China

Springer

Charu C. Aggarwal
IBM
Thomas J. Watson Research
Center
19 Skyline Drive
Hawthorne, NY10532
USA
charu@us.ibm.com

Haixun Wang
Microsoft Research Asia
49 Zhichun Road
100190 Beijing
5F Sigma Center
China, People's Republic
haixunw@microsoft.com

ISSN 1386-2944
ISBN 978-1-4419-6044-3 e-ISBN 978-1-4419-6045-0
DOI 10.1007/978-1-4419-6045-0
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010920842

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

List of Figures

List of Tables

Preface

The field of graph mining has seen a rapid explosion in recent years because of new applications in computational biology, software bug localization, and social and communication networking. This book is designed for studying various applications in the context of managing and mining graphs. Graph mining has been studied by the theoretical community extensively in the context of numerous problems such as graph partitioning, node clustering, matching, and connectivity analysis. However the traditional work in the theoretical community cannot be directly used in practical applications because of the following reasons:

- The definitions of problems such as graph partitioning, matching and dimensionality reduction are too “clean” to be used with real applications. In real applications, the problem may have different variations such as a disk-resident case, a multi-graph case, or other constraints associated with the graphs. In many cases, problems such as frequent sub-graph mining and dense graph mining may have a variety of different flavors for different scenarios.

- The size of the applications in real scenarios are often very large. In such cases, the graphs may not be stored in main memory, but may be available only on disk. A classic example of this is the case of web and social network graphs, which may contain millions of nodes. As a result, it is often necessary to design specialized algorithms which are sensitive to disk access efficiency constraints. In some cases, the entire graph may not be available at one time, but may be available in the form of a continuous stream. This is the case in many applications such as social and telecommunication networks in which edges are received continuously.

The book will study the problem of managing and mining graphs from an applied point of view. It is assumed that the underlying graphs are massive and cannot be held in main memory. This change in assumption has a critical impact on the algorithms which are required to process such graphs. The problems studied in the book include algorithms for frequent pattern mining, graph matching, indexing, classification, clustering, and dense graph mining.In many cases, the problem of graph management and mining has been studied from the perspective of structured and XML data. Where possible, we have clarified the connections with the methods and algorithms designed by the XML data management community. We also provide a detailed discussion of the application of graph mining algorithms in a number of recent applications such as graph privacy, web and social networks.

Many of the graph algorithms are sensitive to the application scenario in which they are encountered. Therefore, we will study the usage of many of these techniques in real scenarios such as the web, social networks, and biological data. This provides a better understanding of how the algorithms in the book apply to different scenarios. Thus, the book provides a comprehensive summary both from an algorithmic and applied perspective.