# Sequence Data Mining

# ADVANCES IN DATABASE SYSTEMS

Series Editor
**Ahmed K. Elmagarmid**

*Purdue University*
*West Lafayette, IN 47907*

**Other books in the Series:**

**DATA STREAMS:** ***Models and Algorithms****, edited by Charu C. Aggarwal*; ISBN: 978- 0-387-28759-1

**SIMILARITY SEARCH**: ***The Metric Space Approach****, P. Zezula,* ***G.*** *Amato, V. Dohnal, M. Batko*; ISBN: 0-387-29146-6

**STREAM DATA MANAGEMENT,** *Nauman Chaudhry, Kevin Shaw, Mahdi Abdelguerfi*; ISBN: 0-387-24393-3

**FUZZY DATABASE MODELING WITH XML,** *Zongmin Ma;* ISBN: 0-387-24248-1

**MINING SEQUENTIAL PATTERNS FROM LARGE DATA SETS**, *Wei Wang and Jiong Yang;* ISBN: 0-387-24246-5

**ADVANCED SIGNATURE INDEXING FOR MULTIMEDIA AND WEB APPLICATIONS,** *Yannis Manolopoulos, Alexandros Nanopoulos, Eleni Tousidou;* ISBN: 1-4020-7425-5

**ADVANCES IN DIGITAL GOVERNMENT*****: Technology, Human Factors, and Policy****, edited by William J. McIver, Jr. and Ahmed K. Elmagarmid;* ISBN: 1-4020-7067-5

**INFORMATION AND DATABASE QUALITY,** *Mario Piattini, Coral Calero and Marcela Genero;* ISBN: 0-7923- 7599-8

**DATA QUALITY,** *Richard Y. Wang, Mostapha Ziad, Yang W. Lee:* ISBN: 0-7923-7215-8

**THE FRACTAL STRUCTURE OF DATA REFERENCE:** ***Applications to the Memory Hierarchy,*** *Bruce McNutt;* ISBN: 0-7923-7945-4

**SEMANTIC MODELS FOR MULTIMEDIA DATABASE SEARCHING AND BROWSING,** *Shu-Ching Chen, R.L. Kashyap, and Arif Ghafoor*; ISBN: 0-7923-7888-1

**INFORMATION BROKERING ACROSS HETEROGENEOUS DIGITAL DATA: A Metadata-based Approach,** *Vipul Kashyap, Amit Sheth*; ISBN: 0-7923-7883-0

**DATA DISSEMINATION IN WIRELESS COMPUTING ENVIRONMENTS,** *Kian-Lee Tan and Beng Chin Ooi*; ISBN: 0-7923-7866-0

**MIDDLEWARE NETWORKS: Concept, Design and Deployment of Internet Infrastructure,** *Michah Lerner, George Vanecek, Nino Vidovic, Dad Vrsalovic;* ISBN: 0-7923-7840-7

**ADVANCED DATABASE INDEXING,** *Yannis Manolopoulos, Yannis Theodoridis, Vassilis J. Tsotras;* ISBN: 0-7923-7716-8

**MULTILEVEL SECURE TRANSACTION PROCESSING,** *Vijay Atluri, Sushil Jajodia, Binto George* ISBN: 0-7923-7702-8

**FUZZY LOGIC IN DATA MODELING,** *Guoqing Chen* ISBN: 0-7923-8253-6

For a complete listing of books in this series, go to http://www.springer.com

# Sequence Data Mining

by

Guozhu Dong
*Wright State University*
*Dayton, Ohio, USA*

and

Jian Pei
*Simon Fraser University*
*Burnaby, BC, Canada*

Springer

Guozhu Dong, PhD, Professor
Department of Computer Science and Eng.
Wright State University
Dayton, Ohio, 45435, USA
e-mail: guozhu.dong@wright.edu

Jian Pei, Ph.D.
Assistant Professor
School of Computing Science
Simon Fraser University
8888 University Drive
Burnaby, BC Canada V5A 1S6
e-mail: jpei@cs.sfu.ca

ISBN-13: 978-0-387-69936-3 e-ISBN-13: 978-0-387-69937-0

Library of Congress Control Number: 2007927815

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

To my parents, my wife and my children. {G.D.}<br>
To my wife Jennifer. {J.P.}

# Foreword

With the rapid development of computer and Internet technology, tremendous amounts of data have been collected in various kinds of applications, and *data mining*, *i.e.*, finding interesting patterns and knowledge from a vast amount of data, has become an imminent task. Among all kinds of data, sequence data has its own unique characteristics and importance, and claims many interesting applications. From customer shopping transactions, to global climate change, from web click streams to biological DNA sequences, the sequence data is ubiquitous and poses its own challenging research issues, calling for dedicated treatment and systematic analysis.

Despite of the existence of a lot of general data mining algorithms and methods, sequence data mining deserves dedicated study and in-depth treatment because of its unique nature of ordering, which leads to many interesting new kinds of knowledge to be discovered, including sequential patterns, motifs, periodic patterns, partially ordered patterns, approximate biological sequence patterns, and so on; and these kinds of patterns will naturally promote the development of new classification, clustering and outlier analysis methods, which in turn call for new, diverse application developments. Therefore, *sequence data mining*, *i.e.*, mining patterns and knowledge from large amount of sequence data, has become one of the most essential and active subfields of data mining research. With many years of active research on sequence data mining by data mining, machine learning, statistical data analysis, and bioinformatics researchers, it is time to present a systematic introduction and comprehensive overview of the state-of-the-art of this interesting theme. This book, by Professors Guozhu Dong and Jian Pei, serves this purpose timely, with remarkable conciseness and in great quality.

There have been many books on the general principles and methodologies of data mining. However, the diversities of data and applications call for dedicated, in-depth, and thorough treatment of each specific kind of data, and for each kind of data, compile a vast array of techniques from multiple disciplines into one comprehensive but concise introduction. Thus there is no wonder to see the recent trend of the publication of a series of new, domain-specific data mining books, such as those on Web data mining, stream data mining, geo-spatial data mining, and multimedia data mining. This book integrates the methodologies of sequence data mining developed in multiple disciplines, including data mining, machine learning, statistics, bioinformatics, genomics, web services, and financial data analysis, into one comprehensive and easily-accessible introduction. It starts with a general overview of the sequence data mining problem, by characterizing the sequence data, sequence patterns and sequence models and their various applications, and then proceeds to different mining algorithms and methodologies. It covers a set of exciting research themes, including sequential pattern mining methods; classification, clustering and feature extraction of sequence data; identification and characterization of sequence motifs; mining partial orders from sequences; distinguishing sequence patterns; and other interesting related topics. The scope of the book is broad, nevertheless the treatment of each chapter is rigorous, in sufficient depth, but still easy to read and comprehend.

Both authors of the book are prominent researchers on sequence data mining and have made important contributions to the progress of this dynamic research field. This ensures that the book is authoritative and reflects the current state of the art. Nevertheless, the book gives a balanced treatment on a wide spectrum of topics, far beyond the authors' own methodologies and research scopes.

Sequence data mining is still a fairly young and dynamic research field. This book may serve researcher and application developers a comprehensive overview of the general concepts, techniques, and applications on sequence data mining and help them explore this exciting field and develop new methods and applications. It may also serve graduate students and other interested readers a general introduction to the state-of-the-art of this promising field.

I find the book is enjoyable to read. I hope you like it too.

*Jiawei Han*
*University of Illinois, Urbana-Champaign*
April 29, 2007

# Biography

Jiawei Han, University of Illinois at Urbana-Champaign

Jiawei Han, Professor, Department of Computer Science, University of Illinois at Urbana-Champaign. His research includes data mining, data warehousing, database systems, data mining from spatiotemporal data, multimedia data, stream and RFID data, Web data, social network data, and biological data, with over 300 journal and conference publications. He has chaired or served on over 100 program committees of international conferences and workshops, including PC co-chair of 2005 (IEEE) International Conference on Data Mining (ICDM).

He is an ACM Fellow and has received 2004 ACM SIGKDD Innovations Award and 2005 IEEE Computer Society Technical Achievement Award. His book "Data Mining: Concepts and Techniques" (2nd ed., Morgan Kaufmann, 2006) has been popularly used as a textbook worldwide.

# Preface

Sequence data is pervasive in our lives. For example, your schedule for any given day is a sequence of your activities. When you read a news story, you are told the development of some events which is also a sequence. If you have investment in companies, you are keen to study the history of those companies' stocks. Deep in your life, you rely on biological sequences including DNA and RNA sequences.

Understanding sequence data is of grand importance. As early as our history can call, our ancestors already started to make predictions or simply conjectures based on their observations of event sequences. For example, a typical task of royal astronomers in ancient China was to make conjectures according to their observations of stellar movements. Even much earlier before that, the nature encodes some "sequence learning algorithms" in lives. For example, some animals such as dogs, mice, and snakes have the capability to predict earthquakes based on environmental change sequences, though the mechanisms are still largely mysteries.

When the general field of data mining emerged in the 1990s, sequence data mining naturally became one of the first class citizens in the field. Much research has been conducted on sequence data mining in the last dozen years. Hundreds if not thousands of research papers have been published in forums of various disciplines, such as data mining, database systems, information retrieval, biology and bioinformatics, industrial engineering, etc. The area of sequence data mining has developed rapidly, producing a diversified array of concepts, techniques and algorithmic tools.

The purpose of this book is to provide, in one place, a concise introduction to the field of sequence data mining, and a fairly comprehensive overview of the essential research results. After an introduction to the basics of sequence data mining, the major topics include (1) mining frequent and closed sequential patterns, (2) clustering, classification, features and distances of sequence data, (3) sequence motifs – identifying and characterizing sequence families, (4) mining partial orders from sequences, (5) mining distinguishing sequence patterns, and (6) overviewing some related topics.

This monograph can be useful to academic researchers and graduate students interested in data mining in general and in sequence data mining in particular, and to scientists and engineers working in fields where sequence data mining is involved, such as bioinformatics, genomics, web services, security, and financial data analysis.

Although sequence data mining is discussed in some general data mining textbooks, as you will see in your reading of our book, we conduct a much deeper and more thorough treatment of sequence data mining, and we draw connections to applications whenever it is possible. Therefore, this manuscript covers much more on sequence data mining than a general data mining textbook.

The area of sequence data mining, although a sub-field of general data mining, is now very rich and it is impossible to cover all of its aspects in this book. Instead, in this book, we tried our best to select several important and fundamental topics, and to provide introductions to the essential concepts and methods, of this rich area.

Sequence data mining is still a fairly young research field. Much more remains to be discovered in this exciting research direction, regarding general concepts, techniques, and applications. We invite you to enjoy the exciting exploration.

## Acknowledgement

Writing a monograph is never easy. We are sincerely grateful to Jiawei Han for his consistent encouragement since the planning stage for this book, as well as writing the foreword for the book. Our deep gratitude also goes to Limsoon Wong and James Bailey for providing very helpful comments on the book. We thank Bin Zhou and Ming Hua for their help in proofreading the draft of this book.

Guozhu Dong is also grateful to Limsoon Wong for introducing him to bioinformatics in the late 1990s. Part of this book was planned and written while he was on sabbatical between 2005 and 2006; he wishes to thank his hosts during this period.

Jian Pei is deeply grateful to Jiawei Han as a mentor for continuous encouragement and support. Jian Pei also thanks his collaborators in the past who have fun together in solving data mining puzzles.

*Guozhu Dong*
*Wright State University*
*Jian Pei*
*Simon Fraser University*
April, 2007

# Contents