Introduction to Concepts and Techniques in Data Mining and Application to Text Mining

Techniques to transform data and information into knowledge with plenty of comprehensible examples

Second Edition

Thanaruk Theeramunkong
Sirindhorn International Institute of Technology
Thammasat University

Sponsored by AIAT.or.th and KINDML, SIIT

About this publication

Title:	:	Introduction to Concepts and Techniques in Data Mining and Application to Text Mining (Techniques to transform data and information into knowledge with plenty of comprehensible examples)
Author	:	Thanaruk Theeramunkong
Cover Designer	:	-
Printing Production	:	Thammasat Printing House (Tha Prachan Campus) Tha Prachan Campus 2 Prachan Road, Phranakorn, Bangkok 10200 Thailand Tel. . +66-2-224-1350, +66-2613-3811, Fax.: +66-2-224-7357 Thammasat Printing House (Rangsit Campus) 99 Paholyothin Rd., Klongluang, Patumtani 12121 Thailand Tel. . +66-2-564-3111 (Ext. 201-204), Fax.: +66-2-564-3119 URL: www.tu.ac.th/org/tuprint/index1.html
Published Date	:	Second Edition: August 2012 First Edition: October 2011 by Technology Promotion Association (Thailand-Japan) (TPA)
Distributor	:	Thammasat University Press (Tha Prachan Campus) Floor U1, Anekprasong 2 building (60th Anniversary building), Prachan Road, Phranakorn, Bangkok 10200 Thailand Tel. . +66-2-223-9232, +66-2623-5110, +66-2613-3801- 2 Fax.: +66-2-226-2083 Thammasat University Press (Rangsit Campus) 2nd Floor, Dome (Administrative) Building, 99 Paholyothin Rd., Klongluang, Patumtani 12121 Thailand Tel. . +66-2-564-4440-59 (Ext. 1204-5), +66-2564--2859-60 Fax.: +66-2-564-2860 URL: www.thammasatpress.com

Theeramunkong, Thanaruk
Introduction to Concepts and Techniques in Data Mining and Application to Text Mining
(Techniques to transform data and information into knowledge with plenty of comprehensible examples

300 p. 1.71 cm. – (2011)
Includes bibliographical references.
ISBN: 978-xxx-xxxx-xx-x

XXX Baht (Thai)
XXX US Dollar

For information, contact thanaruk@siit.tu.ac.th
visit our Web site at *www.siit.tu.ac.th/~thanaruk*
Printed in Thailand
05 06 07 08 09 5 4 3 2 1

Sponsored by AIAT.or.th and KINDML, SIIT

Preface

Recently advances in information technology and communication enable us to store and exchange a large amount of data in both structured form, such as business, scientific, or social events as relational database transactions, and unstructured form, such as textual and multimedia data. Moreover, increasing use of internet and web technology makes us face with a titanic amount of online accessible data and information, and then generates a need to invent an intelligent and practical techniques and tools to reveal meaningful knowledge from such data and information. Known as data mining or knowledge discovery in database, this multidisciplinary field involves database technology, machine learning, pattern recognition, statistics, artificial intelligence, parallel and distributed computing and visualization. Up to present, there have been a large number of published books related to data mining and knowledge discovery. They give either a good introduction or advance/deep knowledge in this field. In contrast with them, this book focuses more on basic concepts and provides many examples in the form of illustrations in order to make readers understand concepts and techniques in data mining easily. The aim of writing this book is to give a comprehensive background for people who have no experience in this field to gain enough knowledge for advance readings.

This book is composed of six chapters. Chapter 1 introduces the field of data mining and text mining. It includes the common steps in data mining and text mining, types and applications of data mining and text mining. Seven types of mining tasks are described and further challenges are discussed. In Chapter 2, data preprocessing is treated in details. It contains how to represent data, how to clean, integrate, transform and reduce data before the main process of data mining. Chapter 3 describes a number of classification and prediction methods, including Fisher's linear discriminant or centroid-based method, k-nearest neighbor method, statistical classifiers, decision trees, rule-based classification, artificial neural networks, and support vector machines. For numeric prediction, linear regression, regression trees and model trees are explained. Moreover, two techniques to use regression as classification are presented. At the end of the chapter, four techniques of model ensemble, namely bagging, boosting, stacking and co-training, are introduced to combine the results from multiple classifiers to obtain better performance. Chapter 4 presents techniques for two general unsupervised learning tasks; cluster analysis and association analysis. For clustering, some common approaches including partition-, hierarchical-, density-, grid-, and model-based clustering, are described in details. Three common algorithms; Apriori, FP-tree and CHARM, are given for association analysis. An extension of association analysis with hierarchical structures is also discussed. Topics of evaluation methods for information retrieval, classification and numeric prediction, forms Chapter 5. Finally, three applications of data mining to text mining are given as examples in Chapter 6. They are centroid-based text classification, document relation extraction and automatic Thai unknown detection. Their original full descriptions can be found in (Lertnattee and Theeramunkong, 2004a), (Sriphaew and Theeramunkong, 2007a) and (TeCho et. al, 2009b).

Finally, the author hopes that this textbook will convey basic concepts on data mining to students and young researchers to get more background for further study in the field. The author would like to thank all members in KINDML (knowledge, Information, Data Management Laboratory) at Sirindhorn International Institute of Technology, Thammasat University for their help and support throughout this book publishing. Verayuth Lertnattee, Kritsada Sriphaew and

Jakkrit TeCho for their contribution to the last chapter, related to text mining application. Thanks to Wittawat Jitkrittum for his help in constructing a number of good examples. Eakasit Pacharawongsakda helps prove-reading and error correction in this book. He, as well as, Nichnan Kittiphattanabawon, Nattapong Tongtep, Nongnuch Ketui, Thawachai Suwanapong help in commenting some parts in the book. Piya Limcharoen designs the cover of this book. Also the author would like to thank all faculty members in KINDML laboratory, including Ekawit Nantajeewarawat, Cholwich Nattee, Pakinee Aimmanee, Surapa Thiemjarus and Boontawee Suntisrivaraporn for their help in several research works in the laboratory which motivates to the writing of this book. Thanks to all KINDML ex-members, including Chutima Pisarn, Ithipan Methasate, Prakasith Kayasith, Thatsanee Charoenporn, Kobkrit Viriyayudhakorn, Swit Phuvipadawat, Arunee Ratikan, Konlakorn Wongpatikaseree, Issariyapol Siriwat, and Peerasak Intarapaiboon for their support. Also thanks to Virach Sornlertlamvanich, Thepchai Supnithi, Choochart Haruechaiyasak, Monthika Boriboon and Krit Kosawat, the NECTEC colleagues who always support our research. Also special thanks to Chutamanee Onsuwan who helps a lot in information extraction project. Finally, many special thanks are given to Taenporn Lertwuthipat from Technology Promotion Association (Thailand-Japan) (TPA, in short), for her great help towards the production of this book.

Thanaruk Theeramunkong
August 2012