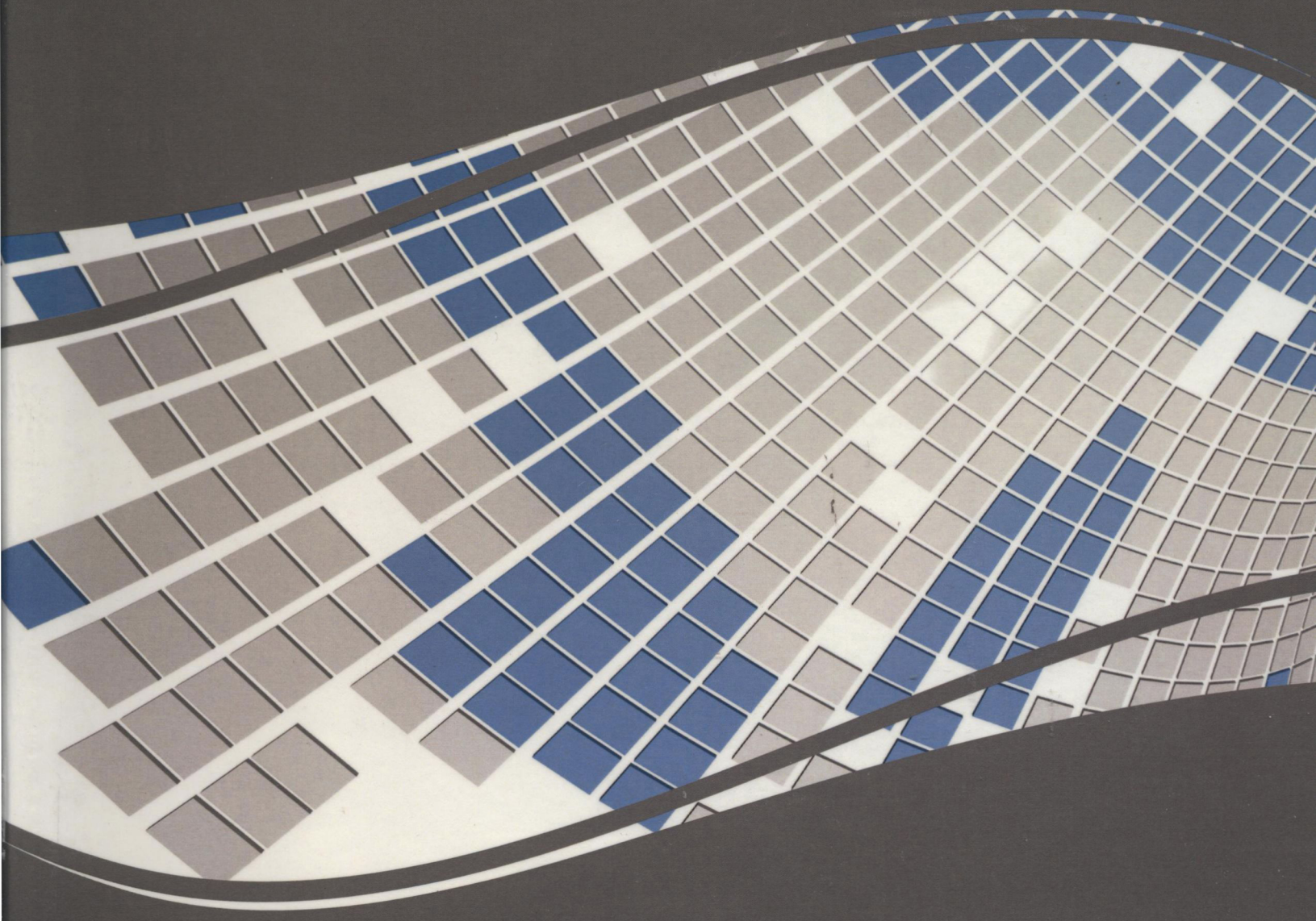
Premier Reference Source

Data Mining and Analysis in the Engineering Field



Vishal Bhatnagar

IGI GLOBAL
DISSEMINATOR OF KNOWLEDGE

Data Mining and Analysis in the Engineering Field

Vishal Bhatnagar
Ambedkar Institute of Advanced Communication Technologies and Research, India

A volume in the Advances in Data Mining and Database Management (ADMDM) Book Series

Information Science REFERENCE
An Imprint of IGI Global

Managing Director: Lindsay Johnston
Production Editor: Jennifer Yoder
Development Editor: Hayley Kang
Acquisitions Editor: Kayla Wolfe
Typesetter: Kaitlyn Kulp
Cover Design: Jason Mull

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

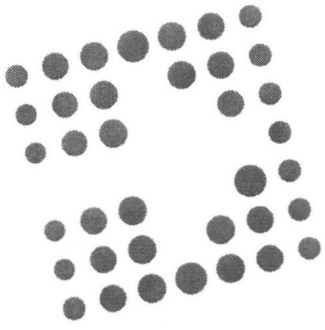
Library of Congress Cataloging-in-Publication Data

Data mining and analysis in the engineering field / Vishal Bhatnagar, editor.
pages cm
Includes bibliographical references and index.
ISBN 978-1-4666-6086-1 (hardcover) -- ISBN 978-1-4666-6087-8 (ebook) -- ISBN 978-1-4666-6089-2 (print & perpetual access) 1. Data mining. I. Bhatnagar, Vishal, 1977-
QA76.9.D343.D372 2014
006.3'12--dc23

2014007990

This book is published in the IGI Global book series Advances in Data Mining and Database Management (ADMDM) (ISSN: 2327-1981; eISSN: 2327-199X)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.



Advances in Data Mining and Database Management (ADMDM) Book Series

David Taniar
Monash University, Australia

ISSN: 2327-1981
EISSN: 2327-199X

Mission

With the large amounts of information available to organizations in today's digital world, there is a need for continual research surrounding emerging methods and tools for collecting, analyzing, and storing data.

The **Advances in Data Mining & Database Management (ADMDM)** series aims to bring together research in information retrieval, data analysis, data warehousing, and related areas in order to become an ideal resource for those working and studying in these fields. IT professionals, software engineers, academicians and upper-level students will find titles within the ADMDM book series particularly useful for staying up-to-date on emerging research, theories, and applications in the fields of data mining and database management.

Coverage

- Cluster Analysis
- Customer Analytics
- Data Mining
- Data Quality
- Data Warehousing
- Database Security
- Database Testing
- Decision Support Systems
- Enterprise Systems
- Text Mining

IGI Global is currently accepting manuscripts for publication within this series. To submit a proposal for a volume in this series, please contact our Acquisition Editors at Acquisitions@igi-global.com or visit: http://www.igi-global.com/publish/.

The Advances in Data Mining and Database Management (ADMDM) Book Series (ISSN 2327-1981) is published by IGI Global, 701 E. Chocolate Avenue, Hershey, PA 17033-1240, USA, www.igi-global.com. This series is composed of titles available for purchase individually; each title is edited to be contextually exclusive from any other title within the series. For pricing and ordering information please visit http://www.igi-global.com/book-series/advances-data-mining-database-management/37146. Postmaster: Send all address changes to above address.

)

Titles in this Series

For a list of additional titles in this series, please visit: www.igi-global.com

Biologically-Inspired Techniques for Knowledge Discovery and Data Mining
Shafiq Alam (University of Auckland, New Zealand)
Information Science Reference • copyright 2014 • 311pp • H/C (ISBN: 9781466660786) • US $265.00 (our price)

Data Mining and Analysis in the Engineering Field
Vishal Bhatnagar (Ambedkar Institute of Advanced Communication Technologies and Research, India)
Information Science Reference • copyright 2014 • 335pp • H/C (ISBN: 9781466660861) • US $225.00 (our price)

Handbook of Research on Cloud Infrastructures for Big Data Analytics
Pethuru Raj (IBM India Pvt Ltd, India) and Ganesh Chandra Deka (Ministry of Labour and Employment, India)
Information Science Reference • copyright 2014 • 570pp • H/C (ISBN: 9781466658646) • US $345.00 (our price)

Innovative Techniques and Applications of Entity Resolution
Hongzhi Wang (Harbin Institute of Technology, China)
Information Science Reference • copyright 2014 • 398pp • H/C (ISBN: 9781466651982) • US $205.00 (our price)

Innovative Document Summarization Techniques Revolutionizing Knowledge Understanding
Alessandro Fiori (IRCC, Institute for Cancer Research and Treatment, Italy)
Information Science Reference • copyright 2014 • 363pp • H/C (ISBN: 9781466650190) • US $175.00 (our price)

Emerging Methods in Predictive Analytics Risk Management and Decision-Making
William H. Hsu (Kansas State University, USA)
Information Science Reference • copyright 2014 • 425pp • H/C (ISBN: 9781466650633) • US $225.00 (our price)

Data Science and Simulation in Transportation Research
Davy Janssens (Hasselt University, Belgium) Ansar-Ul-Haque Yasar (Hasselt University, Belgium) and Luk Knapen (Hasselt University, Belgium)
Information Science Reference • copyright 2014 • 350pp • H/C (ISBN: 9781466649200) • US $175.00 (our price)

Big Data Management, Technologies, and Applications
Wen-Chen Hu (University of North Dakota, USA) and Naima Kaabouch (University of North Dakota, USA)
Information Science Reference • copyright 2014 • 342pp • H/C (ISBN: 9781466646995) • US $175.00 (our price)

Innovative Approaches of Data Visualization and Visual Analytics
Mao Lin Huang (University of Technology, Sydney, Australia) and Weidong Huang (CSIRO, Australia)
Information Science Reference • copyright 2014 • 464pp • H/C (ISBN: 9781466643093) • US $200.00 (our price)

www.igi-global.com

701 E. Chocolate Ave., Hershey, PA 17033
Order online at www.igi-global.com or call 717-533-8845 x100
To place a standing order for titles released in this series, contact: cust@igi-global.com
Mon-Fri 8:00 am - 5:00 pm (est) or fax 24 hours a day 717-533-8661

Editorial Advisory Board

List of Reviewers

Raju Ranjan, *Ideal College of Engineering, India*
Amit Rathi, *Jaypee Institute of Technology, India*
Ibrahem S., *Universiti Teknologi Malaysia, Malaysia*
Pijush Samui, *VIT University, India*
Sanjeev Kumar Sharma, *Oriental Institute of Science and Technology, India*
Sanur Sharma, *GTBIT, India*
Shailendra Kr. Sonkar, *Ashoka Institute of Technology and Mgmt, India*
Tarun Srivastava, *Wipro India Limited, India*
Munish Trivedi, *Dehradun Institute of Technology, India*
Christopher Watson, *Durham University, UK*

Foreword

As digital businesses and society reinvent company processes and lifestyle behaviors, more information-based resources arise outside traditional information technology scenarios. Rapid growth of broadband connections elevates the number of connected people and connectable devices; the exigencies of living with online content and services has led to more than 2.7 billion people being connected to the Internet and 50 billion applications downloaded over all types of devices. Enterprises need to rapidly adjust more than one organizational or production layer to this new fast landscape; it is crucial to establish a novel analytical competitive advantage in order to not succumb to this tsunami of never-ending data generation. One of the main problems is trying to distill knowledge from this incredible source of information made of devices, people, processes, programs, and so on.

Beginning with the research around databases in the late 1980s, the field of data mining has acquired a leading role in the last decade, reinforcing the research community. Ranging from databases to machine learning, computational intelligence, statistics, visualization, and high-performance computing, the need to find synergistic approaches to the complex problem of adding value to data has been approached in various ways.

This book focuses on the analytical techniques of data mining, which are primarily classification, clustering, association rule mining, neural network, and genetic algorithms. These approaches are discussed considering many mathematical prerequisites and several useful application scenarios. The book is accessible to many potential readers, not strictly to those with experience in data mining.

Overall, this edited volume is a very welcome new publication in the arena of data mining, Dr. Vishal Bhatnagar has done an outstanding job by collecting timely and important contributions and opening new scenarios in the research and development of data mining systems.

Vincenzo Loia
University of Salerno, Italy

Vincenzo Loia *received the PhD in Computer Science from the University of Paris VI, France in 1989, and the bachelor degree in Computer Science from the University of Salerno in 1984. Since 1989, he has been a faculty member at the University of Salerno, where he teaches Operating Systems-Based Systems and Multi Agent-Based Systems. His current position is Professor of Computer Science. He was principal investigator in a number of industrial R&D projects and in academic research projects. He is the author of over 300 original research papers in international journals, book chapters, and in international conference proceedings. He has edited 5 research books about agent technology, Internet, and soft computing methodologies. He is Co-Editor-in Chief of Soft Computing, and founder and Editor-in-Chief of Ambient Intelligence and Humanized Computing, both published by Springer-Verlag. He serves as editor in a dozen international journals. His current research interests focus on merging soft computing and agent technology to design technologically complex environments, with particular interest in Web Intelligence and Ambient Intelligence applications.*

Preface

It is all about managing the bulk of the data available in various repositories. Currently, we rarely find a venture that does not have a bulk of data. The situation has come when the flow has gone above 1TB of data storage. The emergence of the data mining tools and techniques has helped in finding nuggets of information from the tera-bytes of data. Data mining is finding hidden and unknown information from large databases. The applications of data mining tools and techniques have grown. Each and every field has come up with uses for data mining tools and techniques. The narrow gap between the use of statistical techniques in data analysis and data mining has resulted in finding mixed applicability. The areas that have gained attention are:

- Software Engineering,
- Civil Engineering,
- Mechanical Engineering,
- Social Network and Mining,
- Web Mining,
- Sentiment Analysis/Opinion Mining,
- Data Labelling,
- Renewable Energy,
- Big Data Analytics,
- Visual Data Mining,
- Use of Data Mining in Fractals.

Data mining tools and techniques are helping to reveal hidden information from the databases. The various statistical and data mining techniques are helping to uncover valuable information from large databases. The major data mining techniques are:

- Classification,
- Clustering,
- Association Rule Mining.

The various statistical techniques are:

- Univariate and Multivariate Data Analysis,
- Hypothesis Testing,

- Use of Mean, Variance, and Standard Deviations in Data Analysis,
- Finding Randomness, etc.

These techniques are able to find widespread application in various engineering fields. The use of the large-scale algorithms has created a revolutionary aspect in the field of data analysis. The complex data analysis is now being eased using the technology of cloud computing, where the computational power of n-processors are used to analyze the data available. The technology of MapReduce and Hadoop has overcome all the hurdles, and now make complex and bulk data computation feasible.

The objective of this publication is to make researchers and other prospective readers aware of the latest trends and patterns in the inclusion of data mining tools and techniques in the engineering areas. The data mining application in the engineering fields of Software Engineering, Fractals, and Virtual Reality is gathered from eminent researchers across the globe. The mission of the publication is to come up with an edited book that discusses the latest and most advanced topic inclusion in contributions from renowned researchers whose work created a revolution in the area. The unique characteristics of the publication are that it:

1. Includes the work of eminent researchers in the application of data mining on emerging engineering areas like Software Engineering, Fractals, and Virtual Reality, which are current topics of research.
2. Is targeted towards providing quality, best, and latest research by eminent researchers considering how it affects common people in their everyday life, like the medical application of fractals or finding new metric and its importance in software engineering.
3. Will influence business users, common people, and society.

The impact of the contribution through this edited book is going to be widespread considering the very fact that areas like software engineering are established and growing fields and finding some significant impact of data mining in such a field will cause a great boom in the software industry, where they are constantly in search of novel and brilliant ways to make effective and productive software. Similarly, finding the data mining viability in the areas of mechanical engineering, which itself is gaining tremendous popularity, will add benefit from finding more and crisp output from the inclusion of data mining. The potential users of this book are:

1. The researchers will be able to know the latest application area of data mining in engineering.
2. The business users will get to know how inclusion of data mining can provide added advantages.
3. The common people will get the secure edge for their data, which is revealed in the social network analysis.

Intended audiences are:

1. Engineers,
2. Researchers,
3. Common People,
4. Business Users.

In my February 2013 call for chapters, I urged and sought contribution to this book from researchers, IT savvy professionals, and young engineers and industrialists across the globe with an aim to extract and accumulate the whole of modern research in the field of application of data mining and analysis using statistical techniques in various engineering fields. The beginning was overwhelming for me as I started to get many chapters with varied applications of the data mining tools and techniques in various fields. The authors whose chapters were selected were asked to include future research directions to enable young engineers and researchers to work in the domain.

The book is a collection of the seventeen chapters by eminent professors, researchers, and industry people from different countries. The chapters were initially peer reviewed by the editorial board members, reviewers, and industry people who themselves span many countries. The chapters are arranged so that all the chapters have the basic introductory topics and the advances as well as future research directions, which enable budding researchers and engineers to pursue their work in this area.

Chapter 1 by P. Vinod, Jikku Kuriakose, T. K. Ansari, and Sonal Ayyappan shows that malware or malicious code intends to harm computer systems without the knowledge of system users. These malicious softwares are unknowingly installed by naive users while browsing the Internet. Once installed, the malware performs unintentional activities like (a) steal username, password; (b) install spy software to provide remote access to the attackers; (c) flood spam messages; (d) perform denial of service attacks; etc. With the emergence of polymorphic and metamorphic malware, signature-based detectors are failing to detect new variants of these malware. The primary reason is that malicious code developed in new generation have different syntactic structures from their predecessor, thereby defeating any pattern matching techniques. Thus, the detection of morphed malware remains a complex open research problem for malware analysts. In this chapter, the authors discuss different types of malware with their detection methods. In addition, they present a proposed method employing machine learning techniques for the detection of metamorphic malware. The methodology demonstrates that appropriately selecting prominent features could improve the classification accuracy. The study also depicts that proposed methods that do not require signatures are effective in identifying and classifying morphed malware.

In chapter 2, Tianxing Cai discusses supply, which is characterized by its diversity, including traditional energy, such as fossil fuels, nuclear power, as well as renewable energy, such as solar, hydroelectric, geothermal, biomass, and wind energy. It involves a complex network system composed of energy generation, energy transformation, energy transportation, and energy consumption. The network does provide the great flexibility for energy transformation and transportation; meanwhile, it presents a complex task for conducting agile energy dispatching when extreme events have caused local energy shortages that need to be restored timely. One of the useful methodologies to solve such a problem is data mining and analysis. Their main objective is to take advantage of inherent tolerance of the imprecision and uncertainty to obtain tractability, robustness, and low solution-cost. The applications and developments of data mining and analysis have amazingly evolved in the last two decades. Many of these applications can be found in the field of renewable energy and energy efficiency where data mining and analysis techniques are showing a great potential to solve the problems that arise in this area. In this chapter, data mining and analysis techniques are briefly introduced. Then the implementation procedures are presented to demonstrate the application of curve fitting for renewable energy network design and optimization, which has the capability to handle the restoration during extreme and emergency situations with uncertain parameters.

Software repositories contain a wealth of information that can be analyzed for knowledge extraction. Software bug repositories are one such repository that stores the information about the defects identified during the development of software as argued by N. K. Nagwani and S. Verma. Information available in software bug repositories like number of bugs priority-wise, component-wise, status-wise, developers-wise, module-wise, summary-terms-wise, can be visualized with the help of two- or three-dimensional graphs. These visualizations help in understanding the bug distribution patterns, software matrices related to the software bugs, and developer information in the bug-fixing process. Visualization techniques are exploited with the help of open source technologies in this chapter to visualize the bug distribution information available in the software bug repositories. Two-dimensional and three-dimensional graphs are generated using java-based open source APIs, namely Jzy3d (Java Easy 3d) and JFreeChart. Android software bug repository is selected for the experimental demonstrations of graphs. The textual bug attribute information is also visualized using frequencies of frequent terms present in it.

Chapter 4 by Naveen Dahiya, Vishal Bhatnagar, Manjeet Singh, and Neeti Sangwan discusses that data mining has proven to be an important technique in terms of efficient information extraction, classification, clustering, and prediction of future trends from a database. The valuable properties of data mining have been put to use in many applications. One such application is Software Development Life Cycle (SDLC), where effective use of data mining techniques has been made by researchers. An exhaustive survey on application of data mining in SDLC has not been done in the past. In this chapter, the authors carry out an in-depth survey of existing literature focused towards application of data mining in SDLC and propose a framework that will classify the work done by various researchers in identification of prominent data mining techniques used in various phases of SDLC and pave the way for future research in the emerging area of data mining in SDLC.

Chapter 5 by Pijush Samui focuses on the determination of pull out capacity (Q) of small ground anchor is an imperative task in civil engineering. This chapter employs three data mining techniques (Genetic Programming [GP], Gaussian Process Regression [GPR], and Minimax Probability Machine Regression [MPMR]) for determination of Q of small ground anchor. Equivalent anchor diameter (D_{eq}), embedment depth (L), average cone resistance (q_c) along the embedment depth, average sleeve friction (f_s) along the embedment depth, and Installation Technique (IT) are used as inputs of the models. The output of models is Q. GP is an evolutionary computing method. The basic idea of GP has been taken from the concept of Genetic Algorithm. GPR is a probabilistic non-parametric modelling approach. It determines the parameter from the given datasets. The output of GPR is a normal distribution. MPMR has been developed based on the principal mimimax probability machine classification. The developed GP, GPR, and MPMR are compared with the Artificial Neural Network (ANN). This chapter also gives a comparative study between GP, GPR, and MPMR models.

Chapter 6 by Manish Kumar and Shashank Srivastava states rules are the smallest building blocks of data mining that produce the evidence for expected outcomes. Many organizations like weather forecasting, production and sales, satellite communications, banks, etc. have adopted this mode of technological understanding not for the enhanced productivity but to attain stability by analyzing past records and preparing a rule-based strategy for the future. Rules may be extracted in different ways depending on the requirements and the dataset from that has to be extracted. This chapter covers various methodologies for extracting such rules. It presents the impact of rule extraction for the predictive analysis in decision making.

Chapter 7 by Jamie Godwin and Peter Matthews explores how labelling of data is an expensive, labour-intensive, and time consuming process and, as such, results in vast quantities of data being unexploited when performing analysis through data mining. This chapter presents a new paradigm using robust multivariate statistical methods to encapsulate normal operational behaviour—not failure behaviour—to autonomously derive unsupervised classifier labels for previously collected data in a rapid, cost-effective manner. This enables traditional machine learning to take place on a much richer dataset. Two case studies are presented in the mechanical engineering domain, namely, a wind turbine gearbox and a rolling element bearing. A statistically sound and robust methodology is contributed, allowing for rapid labelling of data to enable traditional data mining techniques. Model development is detailed, along with a comparative evaluation of the metrics. Robust derivatives are presented and their superiority is shown. Example "R" code is given in the appendix, allowing readers to employ the techniques discussed. High levels of agreement between the derived statistical approaches and the underlying condition of the components can be found, showing the practical nature and benefit of this approach.

Chapter 8 by Gábor Hosszú presents statistical evaluations of script relics. Its concept is exploiting mathematical statistical methods to extract hidden correlations among different script relics. Examining the genealogy of the graphemes of scripts is necessary for exploring the evolution of the writing systems, reading undeciphered inscriptions, and deciphering undeciphered scripts. The chapter focuses on the cluster analysis as one of the most popular mathematical statistical methods. The chapter presents the application of the clustering in the classification of Rovash (pronounced "rove-ash," an alternative spelling: Rovas) relics. The various Rovash scripts were used by nations in the Eurasian Steppe and in the Carpathian Basin. The specialty of the Rovash paleography is that the Rovash script family shows a vital evolution during the last centuries; therefore, it is ideal to test the models of the evolution of the glyphs. The most important Rovash script is the Szekely-Hungarian Rovash. Cluster analysis algorithms are applied for determining the common sets among the significant Szekely-Hungarian Rovash alphabets. The determined Rovash relic ties prove the usefulness of the clustering methods in the Rovash paleography.

Chapter 9 by D. P. Acharjya and Mary A. Geetha explores the fundamental concept of crisp set has been extended in many directions in the recent past. The notion of rough set by Pawlak is noteworthy among them. The rough set philosophy is based on the concept that there is some information associated with each object of the universe. There is a need to classify objects of the universe based on the indiscernibility relation among them. In the view of granular computing, rough set model is researched by single granulation. It has been extended to multigranular rough set model in which the set approximations are defined by using multiple equivalence relations on the universe simultaneously. However, in many real life scenarios, an information system establishes the relation with different universes. This gave the extension of multigranulation rough set on single universal set to multigranulation rough set on two universal sets. This chapter defines multigranulation rough set for two universal sets U and V. In addition, the algebraic properties, measures of uncertainty and topological characterization that are interesting in the theory of multigranular rough sets are studied. This helps in describing and solving real life problems more accurately.

Chapter 10 by Manish Kumar and Sumit Kumar identifies Web usage mining which can extract useful information from Weblogs to discover user access patterns of Web pages. Web usage mining itself can be classified further depending on the kind of usage data. This may consider Web server data, application server data, or application level data. Web server data corresponds to the user logs that are collected at Web servers. Some of the typical data collected at Web server are the URL requested, the IP address from which the request originated, and timestamp. Weblog data is required to be cleaned, condensed, and transformed in order to retrieve and analyze significant and useful information. This chapter analyzes access frequent patterns by applying the FP-growth algorithm, which is further optimized by using Genetic Algorithm (GA) and fuzzy logic.

Chapter 11 by Basant Agarwal and Namita Mittal surveys on Opinion Mining or Sentiment Analysis is the study that analyzes people's opinions or sentiments from the text towards entities such as products and services. It has always been important to know what other people think. With the rapid growth of availability and popularity of online review sites, blogs', forums', and social networking sites' necessity of analysing and understanding these reviews has arisen. The main approaches for sentiment analysis can be categorized into semantic orientation-based approaches, knowledge-based, and machine-learning algorithms. This chapter surveys the machine learning approaches applied to sentiment analysis-based applications. The main emphasis of this chapter is to discuss the research involved in applying machine learning methods mostly for sentiment classification at document level. Machine learning-based approaches work in the following phases, which are discussed in detail in this chapter for sentiment classification: (1) feature extraction, (2) feature weighting schemes, (3) feature selection, and (4) machine-learning methods. This chapter also discusses the standard free benchmark datasets and evaluation methods for sentiment analysis. The authors conclude the chapter with a comparative study of some state-of-the-art methods for sentiment analysis and some possible future research directions in opinion mining and sentiment analysis.

Chapter 12 by Elena Baralis, Luca Cagliero, Saima Jabeen, Alessandro Fiori, and Sajid Shah overviews that with the diffusion of online newspapers and social media, users are becoming capable of retrieving dozens of news articles covering the same topic in a short time. News article summarization is the task of automatically selecting a worthwhile subset of news' sentences that users could easily explore. Promising research directions in this field are the use of semantics-based models (e.g., ontologies and taxonomies) to identify key document topics and the integration of social data analysis to also consider the current user's interests during summary generation. The chapter overviews the most recent research advances in document summarization and presents a novel strategy to combine ontology-based and social knowledge for addressing the problem of generic (not query-based) multi-document summarization of news articles. To identify the most salient news articles' sentences, an ontology-based text analysis is performed during the summarization process. Furthermore, the social content acquired from real Twitter messages is separately analyzed to also consider the current interests of social network users for sentence evaluation. The combination of ontological and social knowledge allows the generation of accurate and easy-to-read news summaries. Moreover, the proposed summarizer performs better than the evaluated competitors on real news articles and Twitter messages.

Chapter 13 by Shailendra Kumar Sonkar, Vishal Bhatnagar, and Rama Krishna Challa argues that dynamic social networks contain vast amounts of data, which is changing continuously. A search in a dynamic social network does not guarantee relevant, filtered, and timely information to the users all the time. There should be some sequential processes to apply some techniques and store the information internally that provides the relevant, filtered, and timely information to the users. In this chapter, the authors categorize the social network users into different age groups and identify the suitable and appropriate parameters, then assign these parameters to the already categorized age groups and propose a layered parameterized framework for intelligent information retrieval in dynamic social network using different techniques of data mining. The primary data mining techniques like clustering group the different groups of social network users based on similarities between key parameter items and by classifying the different classes of social network users based on differences among key parameter items, and it can be association rule mining, which finds the frequent social network users from the available users.

Chapter 14 by Chellammal Surianarayanan and Gopinath Ganapathy argues for implementation of data mining for service-oriented computing. Web services have become the de facto platform for developing enterprise applications using existing interoperable and reusable services that are accessible over networks. Development of any service-based application involves the process of discovering and combining one or more required services (i.e. service discovery) from the available services, which are quite large in number. With the availability of several services, manually discovering required services becomes impractical and time consuming. In applications having composition or dynamic needs, manual discovery even prohibits the usage of services itself. Therefore, effective techniques which extract relevant services from huge service repositories in relatively short intervals of time are crucial. Discovery of service usage patterns and associations/relationships among atomic services would facilitate efficient service composition. Further, with availability of several services, it is more likely to find many matched services for a given query, and hence, efficient methods are required to present the results in useful form to enable the client to choose the best one. Data mining provides well known exploratory techniques to extract relevant and useful information from huge data repositories. In this chapter, an overview of various issues of service discovery and composition and how they can be resolved using data mining methods are presented. Various research works that employ data mining methods for discovery and composition are reviewed and classified. A case study is presented that serves as a proof of concept for how data mining techniques can enhance semantic service discovery.

Chapter 15 by Abdulrahman R. Alazemi and Abdulaziz R. Alazemi explores the advent of information technologies brought with it the availability of huge amounts of data to be utilized by enterprises. Data mining technologies are used to search vast amounts of data for vital insight regarding business. Data mining is used to acquire business intelligence and to acquire hidden knowledge in large databases or the Internet. Business intelligence can find hidden relations, predict future outcomes, and speculate and allocate resources. This uncovered knowledge helps in gaining competitive advantages, better customer relationships, and even fraud detection. In this chapter, the authors describe how data mining is used to achieve business intelligence. Furthermore, they look into some of the challenges in achieving business intelligence.

Chapter 16 by Seyed Jalaleddin Mousavirad and Hossein Ebrahimpour-Komleh presents classification of biomedical data plays a significant role in prediction and diagnosis of disease. The existence of redundant and irrelevant features is one of the major problems in biomedical data classification. Excluding these features can improve the performance of classification algorithm. Feature selection is the problem of selecting a subset of features without reducing the accuracy of the original set of features. These algorithms are divided into three categories: wrapper, filter, and embedded methods. Wrapper methods use the learning algorithm for selection of features while filter methods use statistical characteristics of data. In the embedded methods, feature selection process combines with the learning process. Population-based metaheuristics can be applied for wrapper feature selection. In these algorithms, a population of candidate solutions is created. Then, they try to improve the objective function using some operators. This chapter presents the application of population-based feature selection to deal with issues of high dimensionality in the biomedical data classification. The result shows that population-based feature selection has presented acceptable performance in biomedical data classification.

Chapter 17 by Seyed Jalaleddin Mousavirad and Hossein Ebrahimpour-Komleh shows medical diagnosis is a most important problem in medical data mining. The possible errors of a physician can reduce with the help of data mining techniques. The goal of this chapter is to analyze and compare predictive data mining techniques in the medical diagnosis. To this purpose, various data mining techniques such as decision tree, neural networks, support vector machine, and lazy modelling are considered. Results show data mining techniques can considerably help a physician.

The applications of data mining are so diversified that it cannot be covered in single book. However, with the encouraging research contributed by the researchers in this book, we (contributors), EAB members, and reviewers tried to sum up the latest research domains, development in the business field, and applicable areas. This edited book will serve as a motivating factor for those researchers who have spent years working as data repositories, data analysts, statisticians, and budding researchers, as they will be able to get better response for their collected data irrespective of the domain/field, as data mining is applicable in each and every area. Engineers who are witnessing the never-ending competition in each and every field will get to know that by the application of data mining in their field, a new dimension will be added which will bring success in their field and business as they will replicate the knowledge sharing aspects in various business ventures.

Vishal Bhatnagar
Ambedkar Institute of Advanced Communication Technologies and Research, India