Data Mining For Scientific And Engineering Applications Massive Computing

This review volume provides from both theoretical and application points of views, recent developments and state-of-the-art reviews in various areas of pattern recognition, image processing, machine learning, soft computing, data mining and web intelligence. Machine Interpretation of Patterns: Image Analysis and Data Mining is an essential and invaluable resource for professionals and advanced graduates in computer science, mathematics and life sciences. It can also be considered as an integrated volume to researchers interested in doing interdisciplinary research where computer science is a component.

This book constitutes the proceedings of the 17th International Conference on Discovery Science, DS 2014, held in Bled, Slovenia, in October 2014. The 30 full papers included in this volume were carefully reviewed and selected from 62 submissions. The papers cover topics such as: computational scientific discovery; data mining and knowledge discovery; machine learning and statistical methods; computational creativity; mining scientific data; data and knowledge visualization; knowledge discovery from scientific literature; mining text, unstructured and multimedia data; mining structured and relational data; mining temporal and spatial data; mining data streams; network analysis; discovery informatics; discovery and experimental workflows; knowledge capture and scientific ontologies; data and knowledge integration; logic and philosophy of scientific discovery; and applications of computational methods in various scientific domains.

Commercial applications of data mining in areas such as e-commerce, market-basket analysis, text-mining, and web-mining have taken on a central focus in the JCDD community. However, there is a significant amount of innovative data mining work taking place in the context of scientific and engineering applications that is not well represented in the mainstream KDD conferences. For example, scientific data mining techniques are being developed and applied to diverse fields such as remote sensing, physics, chemistry, biology, astronomy, structural mechanics, computational fluid dynamics etc. In these areas, data mining frequently complements and enhances existing analysis methods based on statistics, exploratory data analysis, and domain-specific approaches. On the surface, it may appear that data from one scientific field, say genomics, is very different from another field, such as physics. However, despite their diversity, there is much that is common across the mining of scientific and engineering data. For example, techniques used to identify objects in images are very similar, regardless of whether the images came from a remote sensing application, a physics experiment, an astronomy observation, or a medical study. Further, with data mining being applied to new types of data, such as mesh data from scientific simulations, there is the opportunity to apply and extend data mining to new scientific domains. This one-day workshop brings together data miners analyzing science data and scientists from diverse fields to share their experiences, learn how techniques developed in one field can be applied in another, and better understand some of the newer techniques being developed in the KDD community. This is the fourth workshop on the topic of Mining Scientific Data sets; for information on earlier workshops, see http://www.ahpcrc.org/conferences/. This workshop continues the tradition of addressing challenging problems in a field where the diversity of applications is matched only by the opportunities that await a practitioner.

The field of data mining is receiving significant attention in today's information-rich society, where data is available from different sources and formats, in large volumes, and no longer constitutes a bottleneck for knowledge acquisition. This rich information has paved the way for novel areas of research, particularly in the crime data analysis realm. Data Mining Trends and Applications in Criminal Science and Investigations presents scientific concepts and frameworks of data mining and analytics implementation and uses across various domains, such as public safety, criminal investigations, intrusion detection, crime scene analysis, and suspect modeling. Exploring the diverse ways that data is revolutionizing the field of criminal science, this publication meets the research needs of law enforcement professionals, data analysts, investigators, researchers, and graduate-level students.

Data mining applications range from commercial to social domains, with novel applications appearing swiftly; for example, within the context of social networks. The expanding application sphere and social reach of advanced data mining raise pertinent issues of privacy and security. Present-day data mining is a progressive multidisciplinary endeavor. This inter- and multidisciplinary approach is well reflected within the field of information systems. The information systems research addresses software and hardware requirements for supporting computationally and data-intensive applications. Furthermore, it encompasses analyzing system and data aspects, and all manual or automated activities. In that respect, research at the interface of information systems and data mining has significant potential to produce actionable knowledge vital for corporate decision-making. The aim of the proposed volume is to provide a balanced treatment of the latest advances and developments in data mining; in particular, exploring synergies at the intersection with information systems. It will serve as a platform for academics and practitioners to highlight their recent achievements and reveal potential opportunities in the field. Thanks to its multidisciplinary nature, the volume is expected to become a vital resource for a broad readership ranging from students, throughout engineers and developers, to researchers and academics.

Data mining aims at finding interesting, useful or profitable information in very large databases. The enormous increase in the size of available scientific and commercial databases (data avalanche) as well as the continuing and exponential growth in performance of present day computers make data mining a very active field. In many cases, the burgeoning volume of data sets has grown so large that it threatens to overwhelm rather than enlighten scientists. Therefore, traditional methods are revised and streamlined, complemented by many new methods to address challenging new problems. Mathematical Programming plays a key role in this endeavor. It helps us to formulate precise objectives (e.g., a clustering criterion or a measure of discrimination) as well as the constraints imposed on the solution (e.g., find a partition, a covering or a hierarchy in clustering). It also provides powerful mathematical tools to build highly performing exact or approximate algorithms. This book is based on lectures presented at the workshop on "Data Mining and Mathematical Programming" (October 10-13, 2006, Montreal) and will be a valuable scientific source of information to faculty, students, and researchers in optimization, data analysis and data mining, as well as people working in computer science, engineering and applied mathematics.

Our ability to generate and collect data has been increasing rapidly. Not only are all of our business, scientific, and government transactions now computerized, but the widespread use of digital cameras, publication tools, and bar codes also generate data. On the collection side, scanned text and image platforms, satellite remote sensing systems, and the World Wide Web have flooded us with a tremendous amount of data. This explosive growth has generated an even more urgent need for new techniques and automated tools that can help us transform this data into useful information and knowledge. Like the first edition, voted the most popular data mining book by KD Nuggets readers, this book explores concepts and techniques for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness, and scalability. However, since the publication of the first edition, great progress has been made in the development of new data mining methods, systems, and applications. This new edition substantially enhances the first edition, and new chapters have been added to address recent developments on mining complex types of data— including stream data, sequence data, graph structured data, social network data, and multi-relational data. A comprehensive, practical look at the concepts and techniques you need to know to get the most out of real business data Updates that incorporate input from readers, changes in the field, and more material on statistics and machine learning Dozens of algorithms and implementation examples, all in easily understood pseudo-code and suitable for use in real-world, large-scale data mining projects Complete classroom support for instructors at www.mkp.com/datamining2e companion site

Mobile communications and ubiquitous computing generate large volumes of data. Mining this data can produce useful knowledge, yet individual privacy is at risk. This book investigates the various scientific and technological issues of mobility data, open problems, and roadmap. The editors manage a research project called GeoPKDD, Geographic Privacy-Aware Knowledge Discovery and Delivery, and this book relates their findings in 13 chapters covering all related subjects.

This book provides an introduction to data science and offers a practical overview of the concepts and techniques that readers need to get the most out of their large-scale data mining projects and research studies. It discusses data-analytical thinking, which is essential to extract useful knowledge and obtain commercial value from the data. Also known as data-driven science, soft computing and data mining disciplines cover a broad interdisciplinary range of scientific methods and processes. The book provides readers with sufficient knowledge to tackle a wide range of issues in complex systems, bringing together the scopes that integrate soft computing and data mining in various combinations of applications and practices, since to thrive in these data-driven ecosystems, researchers, data analysts and practitioners must understand the design choice and options of these approaches. This book helps readers to solve complex benchmark problems and to better appreciate the concepts, tools and techniques used.

This book explains and explores the principal techniques of Data Mining, the automatic extraction of implicit and potentially useful information from data, which is increasingly used in commercial, scientific and other application areas. It focuses on classification, association rule mining and clustering. Each topic is clearly explained, with a focus on algorithms not mathematical formalism, and is illustrated by detailed worked examples. The book is written for readers without a strong background in mathematics or statistics and any formulae used are explained in detail. It can be used as a textbook to support courses at undergraduate or postgraduate levels in a wide range of subjects including Computer Science, Business Studies, Marketing, Artificial Intelligence, Bioinformatics and Forensic Science. As an aid to self-study, it aims to help general readers develop the necessary understanding of what is inside the 'black box' so they can use commercial data mining packages discriminatingly, as well as enabling advanced readers or academic researchers to understand or contribute to future technical advances in the field. Each chapter has practical exercises to enable readers to check their progress. A full glossary of technical terms used is included. Principles of Data Mining includes descriptions of algorithms for classifying streaming data, both stationary data, where the underlying model is fixed, and data that is time-dependent, where the underlying model changes from time to time - a phenomenon known as concept drift. The expanded fourth edition gives a detailed description of a feed-forward neural network with backpropagation and shows how it can be used for classification.

Data Mining is the science and technology of exploring large and complex bodies of data in order to discover useful patterns. It is extremely important because it enables modeling and knowledge extraction from abundant data availability. This book introduces soft computing methods extending the envelope of problems that data mining can solve efficiently. It presents practical soft-computing approaches in data mining and includes various real-world case studies with detailed results.

This book is devoted to current problems of artificial and computational intelligence including decision-making systems. Collecting, analysis, and processing information are the current directions of modern computer science. Development of new modern information and computer technologies for data analysis and processing in various fields of data mining and machine learning creates the conditions for increasing effectiveness of the information processing by both the decrease of time and the increase of accuracy of the data processing. The book contains of 54 science papers which include the results of research concerning the current directions in the fields of data mining, machine learning, and decision making. The papers are divided in terms of their topic into three sections. The first section "Analysis and Modeling of Complex Systems and Processes" contains of 26 papers, and the second section "Theoretical and Applied Aspects of Decision-Making Systems" contains of 13 papers. There are 15 papers in the third section "Computational Intelligence and Inductive Modeling". The book is focused to scientists and developers in the fields of data mining, machine learning and decision-making systems.

This textbook explores the different aspects of data mining from the fundamentals to the complex data types and their applications, capturing the wide diversity of problem domains for data mining issues. It goes beyond the traditional focus on data mining problems to introduce advanced data types such as text, time series, discrete sequences, spatial data, graph data, and social networks. Until now, no single book has addressed all these topics in a comprehensive and integrated way. The chapters of this book fall into one of three categories: Fundamental chapters: Data mining has four main problems, which correspond to clustering, classification, association pattern mining, and outlier analysis. These chapters comprehensively discuss a wide variety of methods for these problems. Domain chapters: These chapters discuss the specific methods used for different domains of data such as text data, time-series data, sequence data, graph data, and spatial data. Application chapters: These chapters study important applications such as stream mining, Web mining, ranking, recommendations, social networks, and privacy preservation. The domain chapters also have an applied flavor. Appropriate for both introductory and advanced data mining courses, Data Mining: The Textbook balances mathematical details and intuition. It contains the necessary mathematical details for professors and researchers, but it is presented in a simple and intuitive style to improve accessibility for students and industrial practitioners (including those with a limited mathematical background). Numerous illustrations, examples, and exercises are included, with an emphasis on semantically interpretable examples. Praise for Data Mining: The Textbook - “As I read through this book, I have already decided to use it in my classes. This is a book written by an outstanding researcher who has made fundamental contributions to data mining, in a way that is both accessible and up to date. The book is complete with theory and practical use cases. It’s a must-have for students and professors alike!" -- Qiang Yang, Chair of Computer Science and Engineering at Hong Kong University of Science and Technology "This is the most amazing and comprehensive text book on data mining. It covers not only the fundamental problems, such as clustering, classification, outliers and frequent patterns, and different data types, including text, time series, sequences, spatial data and graphs, but also various applications, such as recommenders, Web, social network and privacy. It is a great book for graduate students and researchers as well as practitioners." -- Philip S. Yu, UIC Distinguished Professor and Wexler Chair in Information Technology at University of Illinois at Chicago

Good data mining practice for business intelligence (the art of turning raw software into meaningful information) is demonstrated by the many new techniques and developments in the conversion of fresh scientific discovery into widely accessible software solutions. Written as an introduction to the main issues associated with the basics of machine learning and the algorithms used in data mining, this text is suitable foradvanced undergraduates, postgraduates and tutors in a wide area of computer science and technology, as well as researchers looking to adapt various algorithms for particular data mining tasks. A valuable addition to libraries and bookshelves of the many companies who are using the principles of data mining to effectively deliver solid business and industry solutions.

This is the second edition of Wil van der Aalst’s seminal book on process mining, which now discusses the field also in the broader context of data science and big data approaches. It includes several additions and updates, e.g. on inductive mining techniques, the notion of alignments, a considerably expanded section on software tools and a completely new chapter of process mining in the large. It is self-contained, while at the same time covering the entire process-mining spectrum from process discovery to predictive analytics. After a general introduction to data science and process mining in Part I, Part II provides the basics of business process modeling and data mining necessary to understand the remainder of the book. Next, Part III focuses on process discovery as the most important process mining task, while Part IV moves beyond discovering the control flow of processes, highlighting conformance checking, and organizational and time perspectives. Part V offers a guide to successfully applying process mining in practice, including an introduction to the widely used open-source tool ProM and several commercial products. Lastly, Part VI takes a step back, reflecting on the material presented and the key open challenges. Overall, this book provides a comprehensive overview of the state of the art in process mining. It is intended for business process analysts, business consultants, process managers, graduate students, and BPM researchers.

Advances in technology are making massive data sets common in many scientific disciplines, such as astronomy, medical imaging, bio-informatics, combinatorial chemistry, remote sensing, and physics. To find useful information in these data sets, scientists and engineers are turning to data mining techniques. This book is a collection of papers based on the first two in a series of workshops on mining scientific datasets. It illustrates the diversity of problems and application areas that can benefit from data mining, as well as the issues and challenges that differentiate scientific data mining from its commercial counterpart. While the focus of the book is on mining scientific data, the work is of broader interest as many of the techniques can be applied equally well to data arising in business and web applications. Audience: This work would be an excellent text for students and researchers who are familiar with the basic principles of data mining and want to learn more about the application of data mining to their problem in science or engineering.

Data Preparation for Data Mining addresses an issue unfortunately ignored by most authorities on data mining: data preparation. Thanks largely to its perceived difficulty, data preparation has traditionally taken a backseat to the more alluring question of how best to extract meaningful knowledge. But without adequate preparation of your data, the return on the resources invested in mining is certain to be disappointing. Dorian Pyle corrects this imbalance. A twenty-five-year veteran of what has become the data mining industry, Pyle shares his own successful data preparation methodology, offering both a conceptual overview for managers and complete technical details for IT professionals. Apply his techniques and watch your mining efforts pay off-in the form of improved performance, reduced distortion, and more valuable results. On the enclosed CD-ROM, you'll find a suite of programs as C source code and compiled into a command-line-driven toolkit. This code illustrates how the author's techniques can be applied to arrive at an automated preparation solution that works for you. Also included are demonstration versions of three commercial products that help with data preparation, along with sample data with which you can practice and experiment. * Offers in-depth coverage of an essential but largely ignored subject. * Goes far beyond theory, leading you-step by step-through the author's own data preparation techniques. * Provides practical illustrations of the author's methodology using realistic sample data sets. * Includes algorithms you can apply directly to your own project, along with instructions for understanding when automation is possible and when greater intervention is required. * Explains how to identify and correct data problems that may be present in your application. * Prepares miners, helping them head into preparation with a better understanding of data sets and their limitations.

Unstructured text, as one of the most important data forms, plays a crucial role in data-driven decision making in domains ranging from social networking and information retrieval to scientific research and healthcare informatics. In many emerging applications, people's information need from text data is becoming multidimensional—they demand useful insights along multiple aspects from a text corpus. However, acquiring such multidimensional knowledge from massive text data remains a challenging task. This book presents data mining techniques that turn unstructured text data into multidimensional knowledge. We investigate two core questions. (1) How does one identify task-relevant text data with declarative queries in multiple dimensions? (2) How does one distill knowledge from text data in a multidimensional space? To address the above questions, we develop a text cube framework. First, we develop a cube construction module that organizes unstructured data into a cube structure, by discovering latent multidimensional and multi-granular structure from the unstructured text corpus and allocating documents into the structure. Second, we develop a cube exploitation module that models multiple dimensions in the cube space, thereby distilling from user-selected data multidimensional knowledge. Together, these two modules constitute an integrated pipeline: leveraging the cube structure, users can perform multidimensional, multigranular data selection with declarative queries; and with cube exploitation algorithms, users can extract multidimensional patterns from the selected data for decision making. The proposed framework has two distinctive advantages when turning text data into multidimensional knowledge: flexibility and label-efficiency. First, it enables acquiring multidimensional knowledge flexibly, as the cube structure allows users to easily identify task-relevant data along multiple dimensions at varied granularities and further distill multidimensional knowledge. Second, the algorithms for cube construction and exploitation require little supervision; this makes the framework appealing for many applications where labeled data are expensive to obtain.

Data mining is the process of uncovering patterns, associations, anomalies, and statistically significant structures and events in data. It borrows and builds on ideas from many disciplines, ranging from statistics to machine learning, mathematical optimization, and signal and image processing. Data mining techniques are becoming an integral part of scientific endeavors in many application domains, including astronomy, bioinformatics, chemistry, materials science, climate, fusion, and combustion. In this chapter, we provide a brief introduction to the data mining process and some of the algorithms used in extracting information from scientific data sets.

Data mining is the process of extracting hidden patterns from data, and it’s commonly used in business, bioinformatics, counter-terrorism, and, increasingly, in professional sports. First popularized in Michael Lewis’ best-selling Moneyball: The Art of Winning An Unfair Game, it is has become an intrinsic part of all professional sports the world over, from baseball to cricket to soccer. While an industry has developed based on statistical analysis services for any given sport, or even for betting behavior analysis on these sports, no research-level book has considered the subject in any detail until now. Sports Data Mining brings together in one place the state of the art as it concerns an international array of sports: baseball, football, basketball, soccer, greyhound racing are all covered, and the authors (including Hsinchun Chen, one of the most esteemed and well-known experts in data mining in the world) present the latest research, developments, software available, and applications for each sport. They even examine the hidden patterns in gaming and wagering, along with the most common systems for wager analysis.

The main goal of the new field of data mining is the analysis of large and complex datasets. Some very important datasets may be derived from business and industrial activities. This kind of data is known as OC enterprise dataOCO. The common characteristic of such datasets is that the analyst wishes to analyze them for the purpose of designing a more cost-effective strategy for optimizing some type of performance measure, such as reducing production time, improving quality, eliminating wastes, or maximizing profit. Data in this category may describe different scheduling scenarios in a manufacturing environment, quality control of some process, fault diagnosis in the operation of a machine or process, risk analysis when issuing credit to applicants, management of supply chains in a manufacturing system, or data for business related decision-making. Sample Chapter(s). Foreword (37 KB). Chapter 1: Enterprise Data Mining: A Review and Research Directions (655 KB). Contents: Enterprise Data Mining: A Review and Research Directions (T W Liao); Application and Comparison of Classification Techniques in Controlling Credit Risk (L Yu et al.); Predictive Classification with Imbalanced Enterprise Data (S Daskalaki et al.); Data Mining Applications of Process Platform Formation for High Variety Production (J Jiao & L Zhang); Multivariate Control Charts from a Data Mining Perspective (G C Porzio & G Ragozini); Maintenance Planning Using Enterprise Data Mining (L P Khoo et al.); Mining Images of Cell-Based Assays (P Perner); Support Vector Machines and Applications (T B Trafalis & O O Oladunni); A Survey of Manifold-Based Learning Methods (X Huo et al.); and other papers. Readership: Graduate students in engineering, computer science, and business schools; researchers and practioners of data mining with emphazis of enterprise data mining."

This is the first comprehensive book dedicated entirely to the field of decision trees in data mining and covers all aspects of this important technique. Decision trees have become one of the most powerful and popular approaches in knowledge discovery and data mining, the science and technology of exploring large and complex bodies of data in order to discover useful patterns. The area is of great importance because it enables modeling and knowledge extraction from the abundance of data available. Both theoreticians and practitioners are continually seeking techniques to make the process more efficient, cost-effective and accurate. Decision trees, originally implemented in decision theory and statistics, are highly effective tools in other areas such as data mining, text mining, information extraction, machine learning, and pattern recognition. This book invites readers to explore the many benefits in data mining that decision trees offer:: Self-explanatory and easy to follow when compacted; Able to handle a variety of input data: nominal, numeric and textual; Able to process datasets that may have errors or missing values; High predictive performance for a relatively small computational effort; Available in many data mining packages over a variety of platforms; Useful for various tasks, such as classification, regression, clustering and feature selection . Sample Chapter(s). Chapter 1: Introduction to Decision Trees (245 KB). Chapter 6: Advanced Decision Trees (409 KB). Chapter 10: Fuzzy Decision Trees (220 KB). Contents: Introduction to Decision Trees; Growing Decision Trees; Evaluation of Classification Trees; Splitting Criteria; Pruning

Trees; Advanced Decision Trees; Decision Forests; Incremental Learning of Decision Trees; Feature Selection; Fuzzy Decision Trees; Hybridization of Decision Trees with Other Techniques; Sequence Classification Using Decision Trees. Readership: Researchers, graduate and undergraduate students in information systems, engineering, computer science, statistics and management.

Chandrika Kamath describes how techniques from the multi-disciplinary field of data mining can be used to address the modern problem of data overload in science and engineering domains. Starting with a survey of analysis problems in different applications, it identifies the common themes across these domains.

Particularly in the fields of software engineering, virtual reality, and computer science, data mining techniques play a critical role in the success of a variety of projects and endeavors. Understanding the available tools and emerging trends in this field is an important consideration for any organization. Data Mining and Analysis in the Engineering Field explores current research in data mining, including the important trends and patterns and their impact in fields such as software engineering. With a focus on modern techniques as well as past experiences, this vital reference work will be of greatest use to engineers, researchers, and practitioners in scientific-, engineering-, and business-related fields.

Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications brings together all the information, tools and methods a professional will need to efficiently use text mining applications and statistical analysis. Winner of a 2012 PROSE Award in Computing and Information Sciences from the Association of American Publishers, this book presents a comprehensive how-to reference that shows the user how to conduct text mining and statistically analyze results. In addition to providing an in-depth examination of core text mining and link detection tools, methods and operations, the book examines advanced preprocessing techniques, knowledge representation considerations, and visualization approaches. Finally, the book explores current real-world, mission-critical applications of text mining and link detection using real world example tutorials in such varied fields as corporate, finance, business intelligence, genomics research, and counterterrorism activities. The world contains an unimaginably vast amount of digital information which is getting ever vaster ever more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the textual data can be used to unlock new sources of economic value, provide fresh insights into science and hold governments to account. As the Internet expands and our natural capacity to process the unstructured text that it contains diminishes, the value of text mining for information retrieval and search will increase dramatically. Extensive case studies, most in a tutorial format, allow the reader to 'click through' the example using a software program, thus learning to conduct text mining analyses in the most rapid manner of learning possible Numerous examples, tutorials, power points and datasets available via companion website on Elsevierdirect.com Glossary of text mining terms provided in the appendix

With the increasing availability of massive observational and experimental data sets (across a wide variety of scientific disciplines) there is an increasing need to provide scientists with efficient computational tools to explore such data in a systematic manner. For example, techniques such as classification and clustering are now being widely used in astronomy to categorize and organize stellar objects into groups and catalogs, which in turn provide the impetus for scientific hypothesis formation and discovery (e.g., see Fayyad, Djorgovski and Weir (1996); or Cheeseman and Stutz (1996) or Fayyad and Smyth (1999) in a more general context). Data-driven exploration of massive spatio-temporal data sets is an area where there is particular need of data mining techniques. Scientists are overwhelmed by the vast quantities of data which simulations, experiments, and observational instruments can produce. Analysis of spatio-temporal data is inherently challenging, yet most current research in data mining is focused on algorithms based on more traditional feature-vector data representations. Scientists are often not particularly interested in raw grid-level data, but rather in the phenomena and processes which are ''driving'' the data. In particular, they are often interested in the temporal and spatial evolution of specific ''spatially local'' structures of interest, e.g., birth-death processes for vortices and interfaces in fluid-flow simulations and experiments, trajectories of extra-tropical cyclones from sea-level pressure data over the Atlantic and Pacific oceans, and sunspot shape and size evolution over time from daily chromospheric images of the Sun. The ability to automatically detect, cluster, and catalog such objects in principle provides an important ''data reduction front-end'' which can convert 4-d data sets (3 spatial and 1 temporal dimension) on a massive grid to a much more abstract representation of local structures and their evolution. In turn, these higher-level representations provide a general framework and basis for further scientific hypothesis generation and investigation, e.g., investigating correlations between local phenomena (such as storm paths) and global trends (such as temperature changes). In this work we focused on detecting and clustering trajectories of individual objects in massive spatio-temporal data sets. There are two primary technical problems involved. First, the local structures of interest must be detected, characterized, and extracted from the mass of overall data. Second, the evolution (in space and/or time) of these structures needs to be modeled and characterized in a systematic manner if the overall goal of producing a reduced and interpretable description of the data is to be met.

Abstract: This dissertation focuses on the use of citation data to evaluate the impactfulness of research in hydrogeology. This study not only explores research impact, but also applies one of the most useful information technologies: data mining techniques on textual data and a practical hydrogeological problem. Following the Schwartz, Fang and Ibaraki (2002) paper in Ground Water, I examined the citation data from ISI in order to check the stability of the bibliometric data and validation of use of this information. I looked at the citation growth patterns of highly-cited papers from the 80s and used that pattern to predict the citation growth for the highly-cited papers in the next decade. This exercise ensures me the use of citation data and gives us an overview of evolution of science in hydrogeology. "Innovation" of the research is another important key to create its impact besides research topics. Water Resources Research papers from 1991 are selected to compare with papers before and follow-on. The most highly cited papers in 1991 appear to be unique in that there are relatively few papers like them that were published previously. Moreover, these papers were sufficiently influential to produce a relatively large number of similar follow-on papers. However, the citation pattern of some classic papers shows that the activities and impact of follow-on papers gradually decline with time. The results of this study reinforce the importance of being a pioneer in a research strand, strategically shifting research strands, adopting strategies that can facilitate really major research shifts. Applications of data mining techniques on two types of data show the advantage of information technology. I evaluated two general strategies and several variants thereof on the one type of database: textual data. The first strategy is based on Naïve Bayes, a popular text classification algorithm. The second strategy is based on Principle Direction Divisive Partitioning, an unsupervised document clustering algorithm. While the performance of both approaches is quite good, some of the new variants that I examined including one, which involves a combination of these two approaches yield even better results. The other type of database is digital photo images. Statistics information (texture) of digital images (in grayscale) and spatial information along with measured hydraulic conductivities for some area in the outcrop are important attributes in the database. Self Organizing Maps (SOM) clustering with these attributes is applied to cluster small images extracted from the outcrop along with 122 sampling points and successfully predict the hydraulic conductivities for the whole section of the outcrop.

This compendium provides a self-contained introduction to mathematical analysis in the field of machine learning and data mining. The mathematical analysis component of the typical mathematical curriculum for computer science students omits these very important ideas and techniques which are indispensable for approaching specialized area of machine learning centered around optimization such as support vector machines, neural networks, various types of regression, feature selection, and clustering. The book is of special interest to researchers and graduate students who will benefit from these application areas discussed in the book.

Written for drug developers rather than computer scientists, this monograph adopts a systematic approach to mining scientifi c data sources, covering all key steps in rational drug discovery, from compound screening to lead compound selection and personalized medicine. Clearly divided into four sections, the first part discusses the different data sources available, both commercial and non-commercial, while the next section looks at the role and value of data mining in drug discovery. The third part compares the most common applications and strategies for polypharmacology, where data mining can substantially enhance the research effort. The final section of the book is devoted to systems biology approaches for compound testing. Throughout the book, industrial and academic drug discovery strategies are addressed, with contributors coming from both areas, enabling an informed decision on when and which data mining tools to use for one's own drug discovery project.

How can you tap into the wealth of social web data to discover who’s making connections with whom, what they’re talking about, and where they’re located? With this expanded and thoroughly revised edition, you’ll learn how to acquire, analyze, and summarize data from all corners of the social web, including Facebook, Twitter, LinkedIn, Google+, GitHub, email, websites, and blogs. Employ the Natural Language Toolkit, NetworkX, and other scientific computing tools to mine popular social web sites Apply advanced text-mining techniques, such as clustering and TF-IDF, to extract meaning from human language data Bootstrap interest graphs from GitHub by discovering affinities among people, programming languages, and coding projects Build interactive visualizations with D3.js, an extraordinarily flexible HTML5 and JavaScript toolkit Take advantage of more than two-dozen Twitter recipes, presented in O’Reilly’s popular "problem/solution/discussion" cookbook format The example code for this unique data science book is maintained in a public GitHub repository. It’s designed to be easily accessible through a turnkey virtual machine that facilitates interactive learning with an easy-to-use collection of IPython Notebooks.

Handbook of Statistical Analysis and Data Mining Applications, Second Edition, is a comprehensive professional reference book that guides business analysts, scientists, engineers and researchers, both academic and industrial, through all stages of data analysis, model building and implementation. The handbook helps users discern technical and business problems, understand the strengths and weaknesses of modern data mining algorithms and employ the right statistical methods for practical application. This book is an ideal reference for users who want to address massive and complex datasets with novel statistical approaches and be able to objectively evaluate analyses and solutions. It has clear, intuitive explanations of the principles and tools for solving problems using modern analytic techniques and discusses their application to real problems in ways accessible and beneficial to practitioners across several areas—from science and engineering, to medicine, academia and commerce. Includes input by practitioners for practitioners Includes tutorials in numerous fields of study that provide step-by-step instruction on how to use supplied tools to build models Contains practical advice from successful real-world implementations Brings together, in a single resource, all the information a beginner needs to understand the tools and issues in data mining to build successful data mining solutions Features clear, intuitive explanations of novel analytical tools and techniques, and their practical applications

Knowledge Discovery in the Social Sciences helps readers find valid, meaningful, and useful information. It is written for researchers and data analysts as well as students who have no prior experience in statistics or computer science. Suitable for a variety of classes—including upper-division courses for undergraduates, introductory courses for graduate students, and courses in data management and advanced statistical methods—the book guides readers in the application of data mining techniques and illustrates the significance of newly discovered knowledge. Readers will learn to: • appreciate the role of data mining in scientific research • develop an understanding of fundamental concepts of data mining and knowledge discovery • use software to carry out data mining tasks • select and assess appropriate models to ensure findings are valid and meaningful • develop basic skills in data preparation, data mining, model selection, and validation • apply concepts with end-of-chapter exercises and review summaries

Our ability to generate and collect data has been increasing rapidly. Not only are all of our business, scientific, and government transactions now computerized, but the widespread use of digital cameras, publication tools, and bar codes also generate data. On the collection side, scanned text and image platforms, satellite remote sensing systems, and the World Wide Web have flooded us with a tremendous amount of data. This explosive growth has generated an even more urgent need for new techniques and automated tools that can help us transform this data into useful information and knowledge. Like the first edition, voted the most popular data mining book by KD Nuggets readers, this book explores concepts and techniques for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness, and scalability. However, since the publication of the first edition, great progress has been made in the development of new data mining methods, systems, and applications. This new edition substantially enhances the first edition, and new chapters have been added to address recent developments on mining complex types of data- including stream data, sequence data, graph structured data, social network data, and multi-relational data. Whether you are a seasoned professional or a new student of data mining, this book has much to offer you: * A comprehensive, practical look at the concepts and techniques you need to know to get the most out of real business data. * Updates that incorporate input from readers, changes in the field, and more material on statistics and machine learning. * Dozens of algorithms and implementation examples, all in easily understood pseudo-code and suitable for use in real-world, large-scale data mining projects. * Complete classroom support for instructors at www.mkp.com/datamining2e companion site.

A state-of-the-art survey of recent advances in data mining or knowledge discovery. Data mining, or knowledge discovery, has become an indispensable technology for businesses and researchers in many fields. Drawing on work in such areas as statistics, machine learning, pattern recognition, databases, and high performance computing, data mining extracts useful information from the large data sets now available to industry and science. This collection surveys the most recent advances in the field and charts directions for future research. The first part looks at pervasive, distributed, and stream data mining, discussing topics that include distributed data mining algorithms for new application areas, several aspects of next-generation data mining systems and applications, and detection of recurrent patterns in digital media. The second part considers data mining, counter-terrorism, and privacy concerns, examining such topics as biosurveillance, marshalling evidence through data mining, and link discovery. The third part looks at scientific data mining; topics include mining temporally-varying phenomena, data sets using graphs, and spatial data mining. The last part considers web, semantics, and data mining, examining advances in text mining algorithms and software, semantic webs, and other subjects.

This reader-friendly textbook presents a comprehensive review of the essentials of image data mining, and the latest cutting-edge techniques used in the field. The coverage spans all aspects of image analysis and understanding, offering deep insights into areas of feature extraction, machine learning, and image retrieval. The theoretical coverage is supported by practical mathematical models and algorithms, utilizing data from real-world examples and experiments. Topics and features: describes the essential tools for image mining, covering Fourier transforms, Gabor filters, and contemporary wavelet transforms; reviews a varied range of state-of-the-art models, algorithms, and procedures for image mining; emphasizes how to deal with real image data for practical image mining; highlights how such features as color, texture, and shape can be mined or extracted from images for image representation; presents four powerful approaches for classifying image data, namely, Bayesian classification, Support Vector Machines, Neural Networks, and Decision Trees; discusses techniques for indexing, image ranking, and image presentation, along with image database visualization methods; provides self-test exercises with instructions or Matlab code, as well as review summaries at the end of each chapter. This easy-to-follow work illuminates how concepts from fundamental and advanced mathematics can be applied to solve a broad range of image data mining problems encountered by students and researchers of computer science. Students of mathematics and other scientific disciplines will also benefit from the applications and solutions described in the text, together with the hands-on exercises that enable the reader to gain first-hand experience of computing.

Created with the input of a distinguished International Board of the foremost authorities in data mining from academia and industry, The Handbook of Data Mining presents comprehensive coverage of data mining concepts and techniques. Algorithms, methodologies, management issues, and tools are all illustrated through engaging examples and real-world applications to ease understanding of the materials. This book is organized into three parts. Part I presents various data mining methodologies, concepts, and available software tools for each methodology. Part II addresses various issues typically faced in the management of data mining projects and tips on how to maximize outcome utility. Part III features numerous real-world applications of these techniques in a variety of areas, including human performance, geospatial, bioinformatics, on- and off-line customer transaction activity, security-related computer audits, network traffic, text and image, and manufacturing quality. This Handbook is ideal for researchers and developers who want to use data mining techniques to derive scientific inferences where extensive data is available in scattered reports and publications. It is also an excellent resource for graduate-level courses on data mining and decision and expert systems methodology.

Data analysis forms the basis of many modes of research ranging from scientific discoveries to governmental findings. With the advent of machine intelligence and neural networks, extracting and modeling, approaching data has been unimpeachably altered. These changes, seemingly small, affect the way societies organize themselves, deliver services, or interact with each other. Predictive Analysis on Large Data for Actionable Knowledge: Emerging Research and Opportunities provides emerging information on extraction and prediction patterns in data mining along with knowledge discovery. While highlighting the current issues in data extraction, readers will learn new methodologies comprising of different algorithms that automate the multidimensional schema that remove the manual processes. This book is a vital resource for researchers, academics, and those seeking new information on data mining techniques and trends.

This book comprehensively covers the topic of mining biomedical text, images and visual features towards information retrieval. Biomedical and Health Informatics is an emerging field of research at the intersection of information science, computer science, and health care and brings tremendous opportunities and challenges due to easily available and abundant biomedical data for further analysis. The aim of healthcare informatics is to ensure the high-quality, efficient healthcare, better treatment and quality of life by analyzing biomedical and healthcare data including patient's data, electronic health records (EHRs) and lifestyle. Previously it was a common requirement to have a domain expert to develop a model for biomedical or healthcare; however, recent advancements in representation learning algorithms allows us to automatically to develop the model. Biomedical Image Mining, a novel research area, due to its large amount of biomedical images increasingly generates and stores digitally. These images are mainly in the form of computed tomography (CT), X-ray, nuclear medicine imaging (PET, SPECT), magnetic resonance imaging (MRI) and ultrasound. Patients' biomedical images can be digitized using data mining techniques and may help in answering several important and critical questions related to health care. Image mining in medicine can help to uncover new relationships between data and reveal new useful information that can be helpful for doctors in treating their patients.

Mohamed Medhat Gaber “It is not my aim to surprise or shock you – but the simplest way I can summarise is to say that there are now in the world machines that think, that learn and that create. Moreover, their ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be coextensive with the range to which the human mind has been applied” by Herbert A. Simon (1916-2001) 1Overview This book suits both graduate students and researchers with a focus on discovering knowledge from scientific data. The use of computational power for data analysis and knowledge discovery in scientific disciplines has found its roots with the re- lution of high-performance computing systems. Computational science in physics, chemistry, and biology represents the first step towards automation of data analysis tasks. The rational behind the developmentof computationalscience in different - eas was automating mathematical operations performed in those areas. There was no attention paid to the scientific discovery process. Automated Scientific Disc- ery (ASD) [1–3] represents the second natural step. ASD attempted to automate the process of theory discovery supported by studies in philosophy of science and cognitive sciences. Although early research articles have shown great successes, the area has not evolved due to many reasons. The most important reason was the lack of interaction between scientists and the automating systems.