Preface

This book focuses on the relevant research theme of warehousing and mining sensor network data, which is attracting a lot of attention from the Database, Data Warehousing and Data Mining research communities. With this main idea in mind, this book is oriented to fundamentals and theoretical issues of sensor networks as well as sensor network applications, which have become of relevant interest for next-generation intelligent information systems. Sensor network applications are manifolds: from environmental data collection/management to alerting/alarming systems, from intelligent tools for monitoring/managing IP networks to novel RFID-based applications etc.

Sensor network data management poses new challenges that are outside the scope of capabilities of conventional DBMS, where data are represented and managed according to a tuple-oriented approach. As an example, DBMS expose a limited memory that is not compatible with the prominent unbounded-memory requirement of sensor network data, which, ideally, originate an unbounded data flow. In this respect, collecting and querying sensor network data is questioning, and it cannot be accomplished via conventional DBMS-inspired methodologies. Also, time is completely neglected in DBMS, whereas it plays a leading role in sensor network data management.

Under a broader view, sensor network data are a specialized class of data streams, which can be defined as intermittent sources of information. The above-mentioned issues become, in consequence of this, the guidelines for the design and development of next-generation Data Stream Management Systems (DSMS), which can be reasonably intended as the next challenge for data management research. Therefore, under another perspective, warehousing and mining sensor network data, and, more generally, data streams can be viewed as methodologies and techniques on top of DSMS, oriented to extend data-intensive capabilities of such systems. The same happened for conventional DBMS, with OLAP and Data Mining tools.

Warehousing and mining sensor network data research can also be roughly indented as the application of traditional warehousing and mining techniques developed in the context of DBMS for relational data as well as non-conventional data (e.g., textual data, raw data, XML data etc) to novel scenarios drawn by sensor networks. Despite this, models and algorithms developed in conventional data warehousing and mining technologies cannot be applied “as-they-are” to the novel context of sensor network data management, as the former are not suitable to requirements of sensor data, such as: time-oriented processing, multiple-rate arrivals, unbounded memory, single-pass processing etc. From this, it follows the need for designing and developing models and algorithms able to deal with previously-unrecognized characteristics of sensor network intelligent information systems, thus overcoming actual limitations of data warehousing and data mining systems and platforms.

Based on these motivations and pursuing these aims, this book covers a broad range of topics: data warehousing models for sensor network data, intelligent acquisition techniques for sensor network data,
ETL processes over sensor network data, advanced techniques for processing sensor network data, efficient storage solutions for sensor network data, collecting sensor network data, querying sensor network data, query languages for sensor network data, fusion and integration techniques for heterogeneous sensor network data, cleaning techniques over sensor network data, mining sensor network data, frequent item set mining over sensor network data, intelligent mining techniques over sensor network data, mining outliers and deviants over sensor network data, discovery of complex knowledge patterns from sensor network data, privacy preserving issues of warehousing and mining sensor network data etc.

The main mission of this book is represented by the achievement of a high-quality publication on fundamentals, state-of-the-art techniques and future trends of warehousing and mining sensor network data research. Themes proposed by this book are viable since, traditionally, data methodologies play a leading role in the research community. In turn, this is due to the fact that data processing issues are orthogonal issues for a broad range of next-generation systems and applications, among which we recall: distributed databases, data warehouses, data mining tools, information systems, knowledge-based systems etc. In this respect, themes proposed by this book have a plus-value as they are focused on a very interesting application field such as sensor network data management, which can be reasonably considered as one of the most relevant research themes presently.

Therefore, this book expands the sensor networks research field by putting the basis for novel research trends in the context of warehousing and mining sensor network data, via addressing topics that are, at now, rarely investigated, such as data mining query languages for sensor data. Indeed, the most important unique characteristic of this book is represented by its interdisciplinarity across different research fields spanning from traditional DBMS to Data Warehousing and Data Mining, all concerned with the innovative research theme of sensor network data management.

This book consists of fifteen chapters organized in six major sections. The first section, titled “Warehousing and OLAPing Sensor Network Data”, focuses the attention on models, techniques and algorithms for warehousing and OLAPing several kinds of sensor network data, from conventional ones to RFID data, location-based sensor data and streaming mobile object observations. The second section, titled “Mining Sensor Network Data”, moves the attention on several mining sensor network data issues, such as anomaly detection in streaming sensor data, knowledge discovery from sensor network data in order to improve the quality of sensor network comprehension tasks, and outlier detection in wireless sensor networks. The third section, titled “Clustering Sensor Network Data”, is related to clustering techniques and algorithms for sensor network data, with particular emphasis over applications in specialized contexts, such as intelligent acquisition techniques for sensor network data and peer-to-peer data clustering in self-organizing sensor networks. The fourth section, titled “Query Languages and Query Optimization Techniques for Warehousing and Mining Sensor Network Data”, focuses the attention on query methodologies for sensor networks, particularly on intelligent query techniques for sensor network data fusion and optimization approaches for query activities embedded in Data Mining tasks over peer-to-peer sensor networks. The fifth section, titled “Intelligent Techniques for Efficient Sensor Network Data Warehousing and Mining”, moves the attention on intelligent techniques devoted to improve the performance of warehousing and mining sensor network data, such as geographic routing of sensor data in the presence of voids and obstacles, and sensor field resource management approaches aiming at improving Data Mining tasks over sensor networks. Finally, the sixth section, titled “Intelligent Techniques for Advanced Sensor Network Data Warehousing and Mining”, is related to advanced aspects of warehousing and mining sensor network data, such as the synergy between event and stream processing for complex applications, and dynamic security key management schemes in sensor networks.
In the following, chapters of the section “Warehousing and OLAPing Sensor Network Data” are summarized.

In the first chapter, titled “Integrated Intelligence: Separating the Wheat from the Chaff in Sensor Data”, Marcos M. Campos and Boriana L. Milenova investigate the issue of warehousing and analytics of sensor network data, which is an area growing in relevance as more and more sensor data are collected and made available for analysis. Applications that involve processing of streaming sensor data require efficient storage, analysis, and monitoring of data streams. Traditionally, in these applications, RDBMSs have been confined to the storage stage. While contemporary RDBMSs were not designed to handle stream-like data, the tight integration of sophisticated analytic capabilities into the core database engine offers a powerful infrastructure that can more broadly support sensor network applications. Other useful components found in RDBMs include: extraction, transformation and load (ETL), centralized data warehousing, and automated alert capabilities. The combination of these components addresses significant challenges in sensor data applications such as data transformations, feature extraction, mining model build and deployment, distributed model scoring, and alerting/messaging infrastructure. Based on these motivations, chapter “Integrated Intelligence – Separating the Wheat from the Chaff in Sensor Data” discusses the usage of existing RDBMS functionality in the context of sensor network applications.

In the second chapter, titled “Improving OLAP Analysis of Multidimensional Data Streams via Efficient Compression Techniques”, Alfredo Cuzzocrea, Filippo Furfaro, Elio Masciari and Domenico Saccà consider multidimensionality issues of data streams, and propose efficient compression techniques for improving OLAP analysis of multidimensional data streams. Authors state that a relevant problem in dealing with data streams consists in the fact that they are intrinsically multi-level and multidimensional in nature, so that they require to be analyzed by means of a multi-level and a multi-resolution (analysis) model accordingly, like OLAP, beyond traditional solutions provided by primitive SQL-based DBMS interfaces. Despite this, a significant issue in dealing with OLAP is represented by the so-called curse of dimensionality problem, which consists in the fact that, when the number of dimensions of the target data cube increases, multidimensional data cannot be accessed and queried efficiently, due to their enormous size. Starting from this practical evidence, several data cube compression techniques have been proposed during the last years, with alternate fortune. Briefly, the main idea of these techniques consists in computing compressed representations of input data cubes in order to evaluate time-consuming OLAP queries against them, thus obtaining approximate answers. Similarly to static data, approximate query answering techniques can be applied to streaming data, in order to improve OLAP analysis of such kind of data. Unfortunately, the data cube compression computational paradigm gets worse when OLAP aggregations are computed on top of a continuously flooding multidimensional data stream. In order to efficiently deal with the curse of dimensionality problem and achieve high efficiency in processing and querying multidimensional data streams, this chapter proposes novel compression techniques over data stream readings that are materialized for OLAP purposes. This allows us to tame the unbounded nature of streaming data, thus dealing with bounded memory issues exposed by conventional DBMS tools. Overall, chapter “Improving OLAP Analysis of Multidimensional Data Streams via Efficient Compression Techniques” introduces an innovative, complex technique for efficiently supporting OLAP analysis of multidimensional data streams.

In the third chapter, titled “Warehousing RFID and Location-Based Sensor Data”, Hector Gonzalez, Jiawei Han, Hong Cheng and Tianyi Wu focus the attention on the problem of efficiently warehousing RFID and location-based sensor data. Authors recognize that RFID datasets are expected to become commonplace in supply-chain management systems. Warehousing and mining this data is an essential
problem with great potential benefits for inventory management, object tracking, and product procurement processes. Since RFID tags can be used to identify each individual item, enormous amounts of location-tracking data are generated. Furthermore, RFID tags can record sensor information such as temperature or humidity. With such data, object movements can be modeled by movement graphs, where nodes correspond to locations, and edges record the history of item transitions between locations and sensor readings recorded during the transition. In this chapter, benefits of the movement graph model in terms of compact representation, complete recording of spatio-temporal and item level information, and its role in facilitating multidimensional analysis are demonstrated. Compression power, and efficiency in query processing are gained by organizing the model around the concept of gateway nodes, which serve as bridges connecting different regions of graph, and provide a natural partition of item trajectories. Moreover, in chapter “Warehousing RFID and Location-Based Sensor Data” multi-dimensional analysis is provided by a graph-based object movement data cube that is constructed by merging and collapsing nodes and edges according to an application-oriented topological structure.

In the fourth chapter, titled “Warehousing and Mining Streams of Mobile Object Observations”, Salvatore Orlando, Alessandra Raffaetà, Alessandro Roncato and Claudio Silvestri study the problem of warehousing and mining streams of mobile object observations, by discussing how data warehousing technology can be used to store aggregate information about trajectories of mobile objects, and to perform OLAP operations over them. To this end, authors define a data cube with spatial and temporal dimensions, discretized according to a hierarchy of regular grids. Authors analyze some measures of interest related to trajectories, such as the number of distinct trajectories in a cell or, starting from a cell, the distance covered by the trajectories in a cell, the average and maximum speed and the average acceleration of the trajectories in the cell, and the frequent patterns obtained by a data mining process on trajectories. Furthermore, authors focus on some specialized algorithms to transform data, and load the measures in the base cells. Such stored values are used, along with suitable aggregate functions, to compute the roll-up operations. In this case, as authors observe, main issues derive from characteristics of input data, i.e. trajectory observations of mobile objects, which are usually produced at different rates, and arrive in streams in an unpredictable and unbounded way. Finally, chapter “Warehousing and Mining Streams of Mobile Object Observations” also discusses some use cases that would benefit from the proposed framework, in particular in the domain of supervision systems to monitor road traffic (or movements of individuals) in a given geographical area.

In the following, chapters of the section “Mining Sensor Network Data” are summarized.

In the fifth chapter, titled “Anomaly Detection in Streaming Sensor Data”, Alec Pawling, Ping Yan, Julián Candia, Tim Schoenharl and Greg Madey consider a cell phone network as a set of automatically deployed sensors that records movement and interaction patterns of the target population. Authors discuss methods for detecting anomalies in streaming data produced by the cell phone network, and motivate this discussion by describing the Wireless Phone Based Emergency Response (WIPER) system, a proof-of-concept decision support system for emergency response managers. Authors also discuss some of the scientific work enabled by this type of sensor data and the related privacy issues, and describe scientific studies that use the cell phone data set and steps they have taken to ensure the security of the data. Finally, chapter “Anomaly Detection in Streaming Sensor Data” describes the overall decision support system and discusses three methods of anomaly detection that can be applied to the data.

In the sixth chapter, titled “Knowledge Discovery for Sensor Network Comprehension”, Pedro Pereira Rodrigues, João Gama and Luís Lopes explore different characteristics of sensor networks that define new requirements for knowledge discovery, with the common goal of extracting some kind of comprehension
about sensor data and sensor networks, focusing on clustering techniques that provide useful information about sensor networks via representing interactions between sensors. This network comprehension ability is related with sensor data clustering and clustering of the data streams produced by the sensors. A wide range of techniques already exists to assess these interactions in centralized scenarios, but the processing abilities of sensors in distributed algorithms present several benefits that shall be considered in future designs. Also, sensors produce data at high rate. Often, human experts need to inspect these data streams visually in order to decide on some corrective or proactive operations. Therefore, chapter “Knowledge Discovery for Sensor Network Comprehension” asserts that visualization of data streams, and of data mining results, is extremely relevant to sensor data management, and can enhance sensor network comprehension, thus it should be addressed in future works.

In the seventh chapter, titled “Why General Outlier Detection Techniques Do Not Suffice for Wireless Sensor Networks”, Yang Zhang, Nirvana Meratnia and Paul Havinga start from recognizing that raw data collected in wireless sensor networks are often unreliable and inaccurate due to noise, faulty sensors and harsh environmental effects. Sensor data that significantly deviate from normal patterns of sensed data are often called outliers. Outlier detection in wireless sensor networks aims at identifying such readings, which represent either measurement errors or interesting events. Due to numerous shortcomings, commonly-used outlier detection techniques for general data seem not to be directly applicable to outlier detection in wireless sensor networks. In this chapter, authors report on the current state-of-the-art on outlier detection techniques for general data, provide a comprehensive technique-based taxonomy for these techniques, and highlight their characteristics in a comparative view. Furthermore, chapter “Why General Outlier Detection Techniques do not Suffice for Wireless Sensor Networks?” addresses challenges of outlier detection in wireless sensor networks, provides a guideline on requirements that suitable outlier detection techniques for wireless sensor networks should meet, and explains why general outlier detection techniques do not suffice.

In the following, chapters of the section “Clustering Sensor Network Data” are summarized.

In the eighth chapter, titled “Intelligent Acquisition Techniques for Sensor Network Data”, Elena Baralis, Tania Cerquitelli and Vincenzo D’Elia investigate the issue of querying sensor networks, which entails the (frequent) acquisition of appropriate sensor measurements. Since sensors are battery-powered and communication is the main source of power consumption, an important issue in this context is energy saving during data collection. This chapter thoroughly describes different clustering algorithms to efficiently discover spatial and temporal correlation among sensors and sensor readings. Discovered correlations allow the selection of a subset of good quality representatives of the whole network. Rather than directly querying all network nodes, only the representative sensors are queried to reduce the communication, computation and power consumption costs. Finally, chapter “Intelligent Acquisition Techniques for Sensor Network Data” presents several experiments with different clustering algorithms demonstrating the adaptability and the effectiveness of the proposed approach.

In the ninth chapter, titled “Peer-to-Peer Data Clustering in Self-Organizing Sensor Networks”, Stefano Lodi, Gabriele Monti, Gianluca Moro and Claudio Sartori propose and evaluate distributed algorithms for data clustering in self-organizing ad-hoc sensor networks with computational, connectivity, and power constraints. Self-organization is essential in environments with a large number of devices, because the resulting system cannot be configured and maintained by specific human adjustments on its single components. One of the benefits of in-network data clustering algorithms is the capability of the network to transmit only relevant, high level information, namely models, instead of large amounts of raw data, also reducing drastically energy consumption. For instance, a sensor network could directly identify or
anticipate extreme environmental events such as tsunami, tornado or volcanic eruptions notifying only the alarm or its probability, rather than transmitting via satellite each single normal wave motion. In chapter “Peer-To-Peer Data Clustering in Self-Organizing Sensor Networks”, the efficiency and efficacy of the methods is evaluated by simulation measuring network traffic, and comparing the generated models with ideal results returned by density-based clustering algorithms for centralized systems.

In the following, chapters of the section “Query Languages and Query Optimization Techniques for Warehousing and Mining Sensor Network Data” are summarized.

In the tenth chapter, titled “Intelligent Querying Techniques for Sensor Data Fusion”, Shi-Kuo Chang, Gennaro Costagliola, Erland Jungert and Karin Camara focus the attention on sensor network data fusion, which imposes a number of novel requirements on query languages and query processing techniques. A spatial/temporal query language called ΣQL has been proposed by the same authors previously, in order to support the retrieval of multimedia information from multiple sources and databases. In this chapter, authors investigate intelligent querying techniques including fusion techniques, multimedia data transformations, interactive progressive query building and ΣQL query processing techniques using sensor data fusion. Furthermore, chapter “Intelligent Querying Techniques for Sensor Data Fusion” illustrates and discusses tasks and query patterns for information fusion, provides a number of examples of iterative queries and shows the effectiveness of ΣQL in a command-action scenario.

In the eleventh chapter, titled “Query Optimisation for Data Mining in Peer-to-Peer Sensor Networks”, Mark Roantree, Alan F. Smeaton, Noel E. O’Connor, Vincent Andrieu, Nicolas Legeay and Fabrice Camous move the attention on sensor devices, which represent one of the more recent sources of large volumes of generated data in sensor networks where dedicated sensing equipment is used to monitor events and happenings in a wide range of domains, including monitoring human biometrics and behavior. In this chapter, authors propose an approach and an implementation of semi-automated enrichment of raw sensor data, where the sensor data can come from a wide variety of sources. Also, authors extract semantics from the sensor data using the proposed XSENSE processing architecture in a multi-stage analysis. Sensor data values are thus transformed into XML data so that well-established XML querying via XPATH and similar techniques can be followed. In this respect, authors propose to distribute XML data on a peer-to-peer configuration and show, through simulations, what the computational costs of executing queries on this P2P network, will be. Authors validate the proposed approach through the use of an array of sensor data readings taken from a range of biometric sensor devices, fitted to movie-watchers as they watched Hollywood movies. These readings were synchronized with video and audio analysis of the actual movies themselves, where movie highlights were automatically detected, in order to correlate these highlights with observed human reactions. XSENSE architecture is used to semantically enrich both the biometric sensor readings and the outputs of video analysis, into one large sensor database. Chapter “Query Optimisation for Data Mining in Peer-to-Peer Sensor Networks” thus presents and validates a scalable means of semi-automating the semantic enrichment of sensor data, thereby providing a means of large-scale sensor data management which is a necessary step in supporting data mining from sensor networks.

In the following, chapters of the section “Intelligent Techniques for Efficient Sensor Network Data Warehousing and Mining” are summarized.

In the twelfth chapter, titled “Geographic Routing of Sensor Data around Voids and Obstacles”, Sotiris Nikoletseas, Olivier Powell and Jose Rolim start from recognizing that geographic routing is becoming the protocol of choice for many sensor network applications. Some very efficient geographic routing algorithms exist, however they require a preliminary planarization of the communication graph.
Planarization induces overhead which makes this approach not optimal when lightweight protocols are required. On the other hand, georouting algorithms which do not rely on planarization have fairly low success rates and either fail to route messages around all but the simplest obstacles or have a high topology control overhead (e.g. contour detection algorithms). In order to fulfill this gap, chapter “Geographic Routing of Sensor Data around Voids and Obstacles” describes the GRIC algorithm, the first lightweight and efficient on demand (i.e., all-to-all) geographic routing algorithm which does not require planarization, has almost 100% delivery rates (when no obstacles are added), and behaves well in the presence of large communication blocking obstacles.

In the thirteenth chapter, titled “Sensor Field Resource Management for Sensor Network Data Mining”, David J. Yates and Jennifer Xu motivate their research by data mining for wireless sensor network applications. Authors consider applications where data is acquired in real-time, and thus data mining is performed on live streams of data rather than on stored databases. One challenge in supporting such applications is that sensor node power is a precious resource that needs to be managed as such. To conserve energy in the sensor field, authors propose and evaluate several approaches to acquiring, and then caching data in a sensor field data server. Authors show that for true real-time applications, for which response time dictates data quality, policies that emulate cache hits by computing and returning approximate values for sensor data yield a simultaneous quality improvement and cost saving. This “win-win” is because when data acquisition response time is sufficiently important, the decrease in resource consumption and increase in data quality achieved by using approximate values outweighs the negative impact on data accuracy due to the approximation. In contrast, when data accuracy drives quality, a linear trade-off between resource consumption and data accuracy emerges. Authors then identify caching and lookup policies for which the sensor field query rate is bounded when servicing an arbitrary workload of user queries. This upper bound is achieved by having multiple user queries share the cost of a sensor field query. Finally, chapter “Sensor Field Resource Management for Sensor Network Data Mining” discusses the challenges facing sensor network data mining applications in terms of data collection, warehousing, and mining techniques.

In the following, chapters of the section “Intelligent Techniques for Advanced Sensor Network Data Warehousing and Mining” are summarized.

In the fourteenth chapter, titled “Event/Stream Processing for Advanced Applications”, Qingchun Jiang, Raman Adaikkalavan and Sharma Chakravarthly state that event processing in the form of ECA rules has been researched extensively from the situation monitoring viewpoint to detect changes in a timely manner and to take appropriate actions. Several event specification languages and processing models have been developed, analyzed, and implemented. More recently, data stream processing has been receiving a lot of attention to deal with applications that generate large amounts of data in real-time at varying input rates and to compute functions over multiple streams that satisfy quality of service (QoS) requirements. A few systems based on the data stream processing model have been proposed to deal with change detection and situation monitoring. However, current data stream processing models lack the notion of composite event specification and computation, and they cannot be readily combined with event detection and rule specification, which are necessary and important for many applications. In this chapter, authors discuss a couple of representative scenarios that require both stream and event processing, and then summarize the similarities and differences between the event and data stream processing models. The comparison clearly indicates that for most of the applications considered for stream processing, event component is needed and is not currently supported. And conversely, earlier event processing systems assumed primitive (or simple) events triggered by DBMS and other applica-
tions and did not consider computed events. By synthesizing these two and combining their strengths, authors present an integrated model – one that will be better than the sum of its parts. Authors discuss the notion of a semantic window, which extends the current window concept for continuous queries, and stream modifiers in order to extend current stream computation model for complicated change detection. Authors further discuss the extension of event specification to include continuous queries. Finally, chapter “Event/Stream Processing for Advanced Applications” demonstrates how one of the scenarios discussed earlier can be elegantly and effectively modeled using the integrated approach.

Finally, in the fifteenth chapter, titled “A Survey of Dynamic Key Management Schemes in Sensor Networks”, Biswajit Panja and Sanjay Kumar Madria observe that, in sensor networks, the large numbers of tiny sensor nodes communicate remotely or locally among themselves to accomplish a wide range of applications. However, these networks pose serious security protocol design challenges due to ad hoc nature of the communication and the presence of constraints such as limited energy, slower processor speed and small memory size. To secure such a wireless network, efficient key management techniques are important as existing techniques from mobile ad hoc networks assume resource-equipped nodes. There are some recent security protocols that have been proposed for sensor networks and some of them have also been implemented in a real environment. This chapter provides an overview of research in the area of key management for sensor networks mainly focused on using a cluster head based architecture. First, authors provide a review of the existing security protocols based on private/public key cryptography, Kerberos, Digital signatures and IP security. Next, authors investigate some of the existing work on key management protocols for sensor networks along with their advantages and disadvantages. Finally, chapter “A Survey of Dynamic Key Management Schemes in Sensor Networks” explores some new approaches for providing key management, cluster head security and dynamic key computations.

Overall, this book represents a solid research contribution to state-of-the-art studies and practical achievements in warehousing and mining sensor network data, and puts the basis for further efforts in this challenging scientific field that will more and more play a leading role in next-generation Database, Data Warehousing and Data Mining research.

Alfredo Cuzzocrea
Editor