uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Alexandre Shema Habyalimana
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.A.Sc. (Electrical and Computer Engineering)
GRADE / DEGREE

School of Information Technology and Engineering
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Target-Based Association Rules for Point-of-Coverage Wireless Sensor Network

TITRE DE LA THÈSE / TITLE OF THESIS

A. Boucherche
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

A. El-Saddik

G. Wainer

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

TARGET-BASED ASSOCIATION RULES FOR POINT-OF-COVERAGE WIRELESS SENSOR NETWORK

by

Alexandre Shema Habyalimana

A thesis submitted to the Faculty of Graduate and Postdoctoral Studies

in partial fulfillment of the requirements for the degree of

MASTER OF APPLIED SCIENCE

in Electrical Engineering

Ottawa-Carleton School of Information Technology and Engineering

University of Ottawa

Ottawa, Canada

January 2010

Your file Votre référence
ISBN: 978-0-494-74174-0
Our file Notre référence
ISBN: 978-0-494-74174-0

Abstract

Wireless sensor networks allow remote and accurate monitoring of physical environment, but these networks are often deployed in inhospitable and nearly inaccessible environments. The lack of accessibility due to harsh environments makes routine maintenance of wireless sensor networks impossible. The networks must then be able to maintain themselves unattended. These accessibility and environmental constraints make resource management an essential requirement of any wireless sensor network in order to ensure continued performance and maintain quality of service to acceptable levels. Intelligent use of resources of wireless sensor networks could be achieved by using modern techniques of knowledge discovery and data mining. This thesis considers a special type of wireless sensor networks named Point-of-Coverage Wireless Sensor Networks. The Thesis proposes the Target-based Association Rules, and associated data preparation mechanisms to improve the performance and quality of service of Point-of-Coverage wireless sensor networks by reducing the energy consumption of the knowledge discovery process. To implement our solution, we exploit the nature of Point-of-Coverage wireless sensor networks and divide sensor nodes into disjoint groups that independently fully cover all of the targets nodes in the field of interest. Data gathering is then performed by each group, alternatively, using the proposed data preparation mechanisms. Then, the gathered data are used by the sink to generate the Target-Based Association rules. Simulation of the proposed data preparation mechanisms shows the benefit of including energy consideration when data mining wireless sensor networks.

Acknowledgements

I want to thank my thesis Supervisor Professor Azzedine Boukerche, and Dr. Samer Samarah for their insight and suggestions. I also thank the students in the electrical and computer department who provided many useful tips and comments.

I also acknowledge my family members for their sacrifices and insisting encouragement throughout my school life. I dedicate this work to my mother Agnès Mukamurenzi, my father Emmanuel Habyarimana, my sister Elisabeth Mukamana who read and proof read the manuscripts, and my brothers Jean Luc Bizimana, Paul Antoinne Ngabonziza, Serge Majyera, and Gigi.

I am unable to finish without adequately expressing my debts to my friends in Ottawa, especially Emile Niyonzima and Antoine Ndayirukiye, for their continued support. Last but not least, my gratitude also goes to all of whom supported me as I went forward, providing comfort and support.

In spite of all the help, support, encouragement and constructive criticism offered by these and other people, I alone, of course, remain responsible for the books shortcoming, faults, and failings.

Contents

List of Tables

List of Figures

Chapter 1

THESIS INTRODUCTION

1.1 Introduction

Wireless sensor networks, hereafter abbreviated WSN, are often deployed in inhospitable and nearly inaccessible environments for remote monitoring. These environmental and accessibility constraints make fault management and resource management essential requirements of any WSN. This thesis provides a solution to improve the Quality of Service (QoS) of WSN using data mining techniques. The solution comprises of two elements. The first element defines association rules in the context of Point-of-Coverage WSN, hereafter abbreviated POCW. The second element describes how to prepare the data that will be used to extract these association rules. The solution provides mechanisms that reduce energy consumption of WSN during the data preparation stage of knowledge discovery process.

The rest of this chapter is organized as follows. Section 1.1 describes what the thesis is all about. Section 1.2 summarizes the motivating research question we are

trying to answer. Section 1.3 provides a concise statement of the question that this thesis tackles. Section 1.4 summarizes the proposed answer and an overview of its performance. Finally, Section 1.5 concludes the chapter.

1.2 Motivations

WSN are made of a large number of small, battery-powered, memory-constraint devices named *sensor nodes*. Each sensor node is capable of local processing: data acquisition and processing, storage and wireless communication. The sensor nodes collaborate among themselves to establish a sensing network wherein sensor nodes uses that sensing network to convey data [1].

However, the wide use of WSN is limited by several challenges that must be overcome. These challenges stem from the physical nature of devices that constitute WSN in general. The used devices are resources limited and this impedes WSN performance which in turn affect the QoS and makes WSN unreliable [1, 2]. Several solutions, such as clustering, multi-hop transmission, data aggregation and fusion, and knowledge discovery techniques [3, 4, 5, 6, 7] have been proposed to cope with these limitations.

In the context of WSN, the use of Knowledge Discovery (KD) is relatively new. The aim of using KD is to provide useful solutions that can be used to cope with some of the challenges outlined in the previous paragraph. KD techniques are used to identify patterns between sensor nodes in WSN. To dig out these patterns, KD relies on data gathered by sensor nodes during their operational time. These data are then mined to extract what is called behavioral patterns. The patterns capture

the relationships between the sensor nodes or their readings. Behavioral patterns can provide knowledge that may be used in decision making in order to improve the performance and the Quality of Services of a WSN [3, 8, 9].

The definition of Knowledge discovery in WSN continues to be improved. In general, the definition is similar to those introduced in the traditional database systems, with slight modifications that reflect the inherent nature and limitations of WSN.

Association Rules are among the first KD techniques proposed to find interesting relationships between sensor nodes in a WSN [3]. Initially, the rules, named Sensor Association Rules, were applied to WSN to find temporal relationships or associations between sensor nodes [3]. These rules are important when predicting possible source of future events and identifying sets of temporally correlated sensor nodes.

However, the sensor association rules are not appropriate when applied to special topologies like the one shown on figure 1.1. This topology consists of sensor nodes randomly deployed around a set of targets at fixed locations. This topology is referred to by Point-of-Coverage Wireless Sensor Network (POCW). POCW is found mostly in military applications such as battlefield's targets monitoring, or in environmental applications such as contaminant flow control. We will describe POCW topology in details in the background chapter.

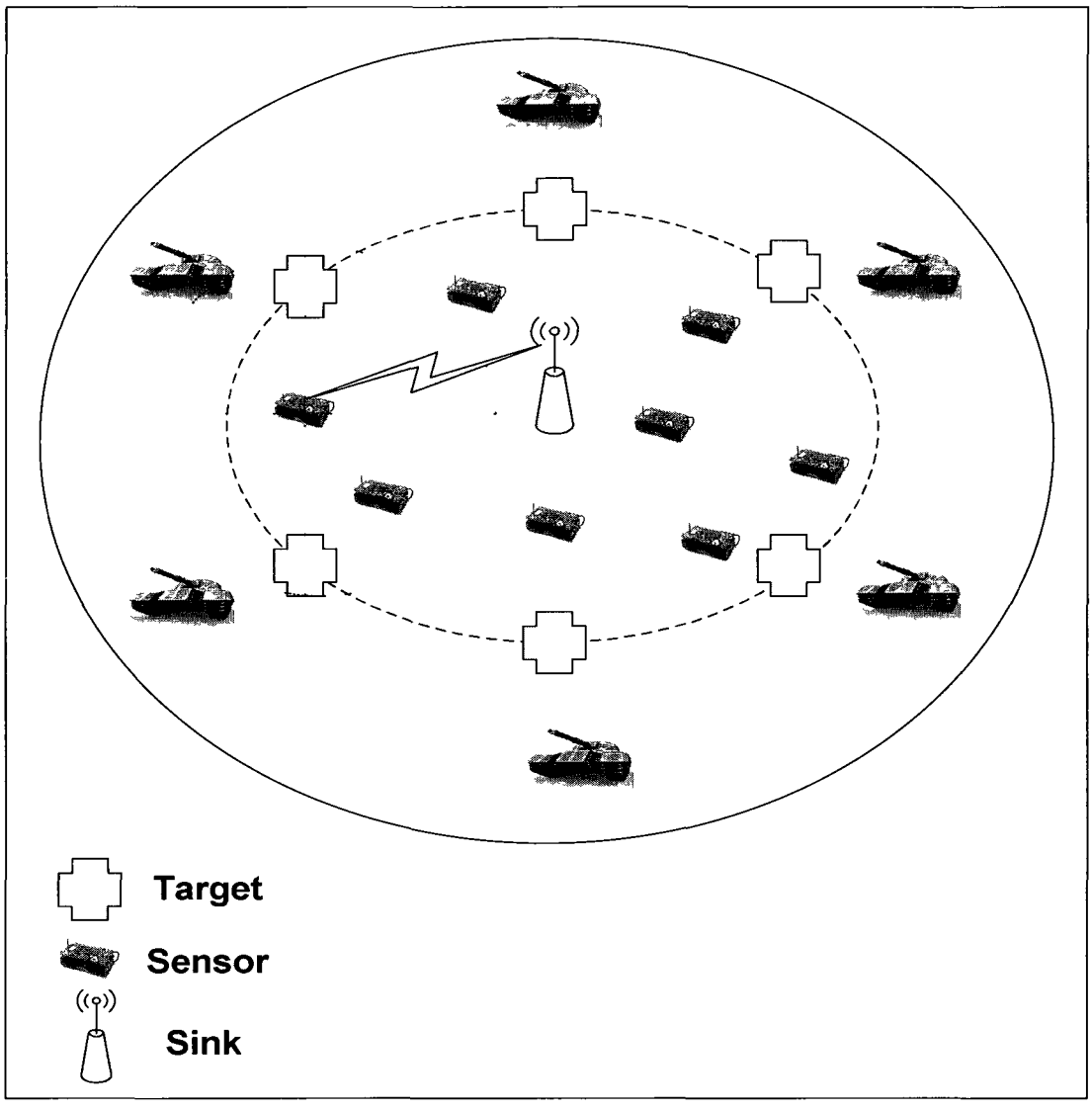

Figure 1.1: Point-of-Coverage Wireless Sensor Network in Battlefield Surveillance

1.3 Research Problem Statement

1.3.1 Research Problem Definition

The objective of this thesis is: first, to define a new kind of behavioral patterns for

point-of-coverage wireless sensor networks; second, to find energy aware mechanisms to be used during the preparation of data needed for generating the proposed behavioral patterns. To achieve this second objective, our focus is on the *Data Preparation* step of the first phase of KDD process. Data preparation is the collection, cleaning, and transformation of data before the processing phase of the knowledge discovery process. We will assume that all the other steps of the first phase have been completed (see subsection 2.2.3.1). As will be explained later on, in subsection 2.2.3.1, those other steps seek to understand the WSN objectives and to translate those objectives into a knowledge discovery problem and also to identifies the data requirements to achieve these objectives.

1.3.2 Problem Justification

As mentioned above, this thesis will define a new kind of behavioral patterns for POCW and the associated data preparation mechanisms for knowledge discovery. Although the importance of data preparation in the KDD process is recognized in WSN [10, 11] and business applications [12, 13], there are no associations rules specific to POCW and no consideration is given to the energy consumption associated to data preparation in these specific networks. This thesis is an attempt to fill this gap by building on recent advances in data mining application in WSN: [14, 15, 3, 8, 16, 17, 18]. The main goal is to improve the energy efficiency of the KDD process and the POCW performance.

1.3.3 Why answer this Problem

We know that communications consume most energy in the context of WSN and we want to limit these costs as much as possible. Of the entire KDD process, one particular phase drain most energy at the sensor node level i.e. the transmission of sensed data prior to data processing (data mining) during data preparation in the pre-processing phase. This is why energy considerations could help extend the network lifetime by reducing energy consumption associated with data preparation.

In the following chapters, we will considered three different approaches to energy efficiency in data preparation. We will begin by describing these approaches then provide their performance evaluation using simulations.

1.4 Proposed Solution Overview

Sensor associations rules provide the means to enhance both the performance and the Quality of Service of WSN. Such association rules can help to predict interesting patterns. However; in some applications, for example in POCW, the interest is in the capture of patterns between targets instead of sensor nodes. We, therefore, introduce a slightly modified version of Sensor Association Rules scheme that reflects the nature of the POCW. We refer to this new technique by Target-based Association Rules. In contrast to Sensor Association Rules, Target-based Rules discover the correlation between the set of targets covered by the POCW network. The patterns are extracted from the targets' activity records transmitted by sensor nodes.

We also introduce, Three mechanisms for preparing the data used to generate these

Target-based association Rules. The mechanisms are named All-Nodes, Schedule-Buffer and Fused-Schedule-Buffer data preparation mechanisms.

1.5 Summary

We will present new behavioral patterns for POCW and its associated data preparation mechanisms. Several simulation experiments will be conducted to evaluate the performance of the proposed data preparation mechanisms and the results will be presented in this thesis.

Chapter 2

RELATED WORK

This chapter introduces the technological background used throughout the thesis development. The technologies in questions are wireless sensor networks, more specifically the Point-of-Coverage wireless sensor networks, and the techniques used for knowledge discovery in database.

This chapter is organized into three main divisions. The first division is an introduction to wireless sensor networks in general. The second division introduces the process of knowledge discovery in databases. The last division provides a review of the state of the art in Association Rules and data preparation mechanisms. The chapter is thus organized as follows. Section 2.1 describes wireless sensor networks in general and Point-of-Coverage wireless sensor networks in particular. The section also defines the concept of observation and event, and outlines the scalability requirements in the context of wireless sensor networks. Section 2.2 provides a brief background of the process of knowledge discovery in databases. Finally, section 2.3 presents the state of the art research in the application of knowledge discovery in database in wireless

sensor networks.

2.1 Wireless Sensor Networks

A wireless sensor network, hereafter abbreviated WSN, is a wireless network consisting of spatially distributed autonomous devices using sensors to cooperatively monitor physical or environmental phenomenon [1]. WSN are usually deployed in inaccessible area either inside the monitored phenomenon or very close to it.

As illustrated in figure 2.1, WSN are made of a large number of small, battery-powered, memory-constraint devices named *sensor nodes*. Each sensor node is capable of local processing (data acquisition and processing), data storage and wireless communications. The sensor nodes collaborate among themselves to establish a wireless sensing network. Then, each sensor node uses this sensing network to convey the gathered data [1] to the sink or the network's user. A wireless sensor network can then be described as a collection of sensor nodes which collaborate with one another to perform a specific function [19].

WSN provide access to information, anytime, anywhere, by: collecting, processing, analyzing and disseminating data. Enabling simultaneously: information gathering, information processing and reliable monitoring for a variety of environment [20]. Reliable in the sense that the combination of the density of sensor nodes and their distributed nature allow large coverage and a high level of redundancy. The result is that vast quantities of sensing information are gathered. Once this information is processed and aggregated, it presents a multidimensional view of the environment under monitoring. As a result, WSN have numerous potential applications. Exam-

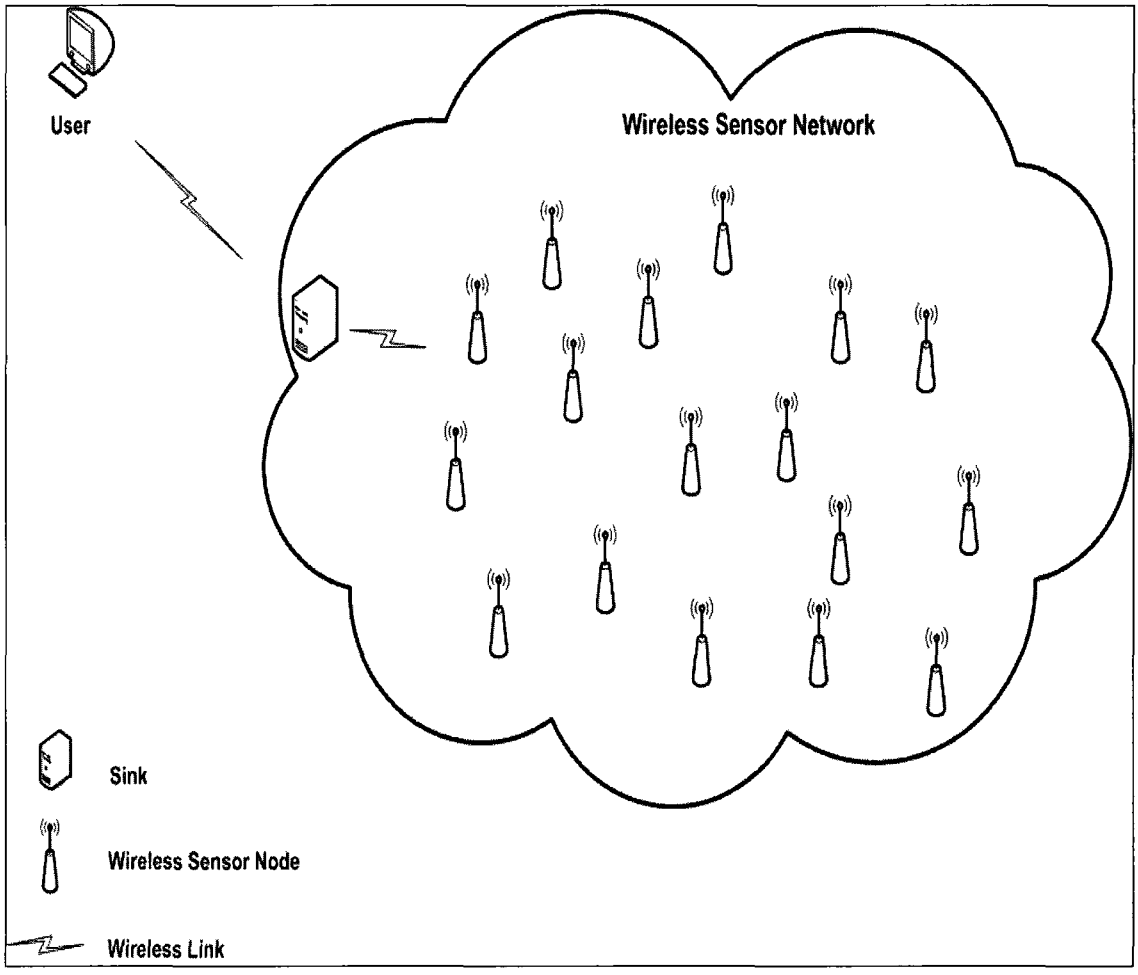

Figure 2.1: A Wireless Sensor Network can be described as a collection of sensor nodes which collaborate with one another to perform a specific function

ples includes environmental monitoring: habitat monitoring, air monitoring, soil and water monitoring, seismic detection, smart homes and target tracking in military surveillance as shown in figure 2.2.

The large volume of data provided by WSN create the need for efficient data handling techniques (processing, transmission, storage) to allow the WSN to achieve its functions [19] and extend its lifetime. Unfortunately, the size of constituting sensor nodes severely limit the sensor node onboard resources and by extension the resources

of the overall network. Moreover, once a WSN is deployed, the hostile environment limit or prevent access to individual sensor nodes for maintenance purposes and the shear number of sensor nodes involved makes repair unpractical in reality. In addition to the maintenance issue, sensor nodes communications consume most of the energy [21]. For a resource limited devices like sensor nodes, energy becomes the scarcest resource of all the resources and it determines the lifetimes of the entire WSN.

Despite all previously mentioned factor that may afflict any WSN, their deployment is relatively simple and inexpensive when compared to traditional wired networks. In addition, WSN can be scaled in size simply by adding more sensor nodes without any complex reconfiguration. However, given the low cost of each sensor node and its individual lifetime which only stretch to a few years ideally, it is far easier and convenient to discard a dead sensor node by replacing it with a new one because it becomes impractical or impossible for example to replace the batteries of each sensor node given the large number that may be deployed, not to mention the limited accessibility after deployment. Hence, operating a WSN requires and relies on energy efficient protocols to reduce the power consumption in order to attain and possibly exceed the planed lifetime of each wireless sensor nodes and by extension to increase the overall lifetime of the WSN [22]. Needless to say that effective use of WSN requires scalability and self-organizing capability [19] as well. Despite the above major constraint, the advantages of WSN, especially their self-reconfiguration capability in response to changing environmental conditions or in response to evolution in the network functions, make their use worthwhile.

At this point, in order to avoid any ambiguity in the discussion to follow, we

shall first define the following terms: *Coverage Area, Sensing Area, Network Field of interest, Sponsor Set*, other definition will be given when the necessity arises.

Coverage Area or **Sensing Area** is the area within the sensing range of a sensor node or of the entire wireless sensor network.

Network Field of Interest is the area or spatial distribution of the phenomenon being monitored.

Sponsor Sets are disjoint sets or groups of sensor nodes that can completely and reliably cover the field of interest or all the targets [8].

Overlapping of different sensor nodes coverage area may result from the random spatial distribution of sensor nodes in the network field of interest. If we consider a case where there is continuous sensing by a sensor nodes, there is a potential for improvement in terms of energy consumption if the network configuration is done right. Therefore, the lifetime of WSN can be increased by simply taking advantage of this unavoidable overlap in sensor nodes sensing area to improve the energy consumptions of sensor nodes. This is attempted by dividing all sensor nodes of the WSN into sponsor sets based on neighboring sensor nodes coverage area. At any given time only one sponsor set is activated to perform all the sensing operations while all the remaining sponsor sets are put in sleep mode until the active sponsor set reach its scheduled time (ideally at the end of its estimated lifetime). Hand off of sensing duty between different sponsor set is determined by the sink during scheduling, and sponsor set are assumed to become active when their scheduled time is reached. Thus, no additional algorithm is used for waking up sponsor set when their scheduled time comes.

2.1.1 Point-of-Coverage WSN

WSN designs in the literature differ greatly in their characteristics and intended use. In this thesis, we consider a large-scale WSN with sensor nodes spread out over an area whose approximate geographic boundaries are known to the network operators.

The nature of the application determines the topology of the wireless sensor network and the way in which the sensor nodes and the Sink node interact. In this section, we describe an architecture we refer to by Point-of-Coverage Wireless Sensor Network, thereafter abbreviated POCW.

Formally, Point-of-Coverage Wireless Network consists of a set of targets $\{T_1, T_2,...., T_n\}$, a set of sensor nodes $\{s_1, s_2,...., s_m\}$, and a Sink node. Sensor nodes are deployed within a field of interest and each sensor node will be responsible for monitoring a set of targets located within its sensing range. Figure 2.2 shows an example of POCW. It may happen that a certain sensor node covers several targets, as illustrated in the figure. Having many sensor nodes covering the same target, and a particular sensor node covering many targets is one of the required properties of POCW.

The traditional operational mode of POCW is to activate all the sensor nodes, in the network, at the same time. Once an event is detected at a particular target, a notification message from the sensor node(s) covering that target is sent to the sink node. Although this mode of interactions sounds simple and easy, it is a costly solution in terms of energy consumption, because redundant data are reported to the sink when coverage areas overlap i.e. several sensor nodes cover the same target.

One way to reduce energy consumption in the POCW is then to use sponsor sets.

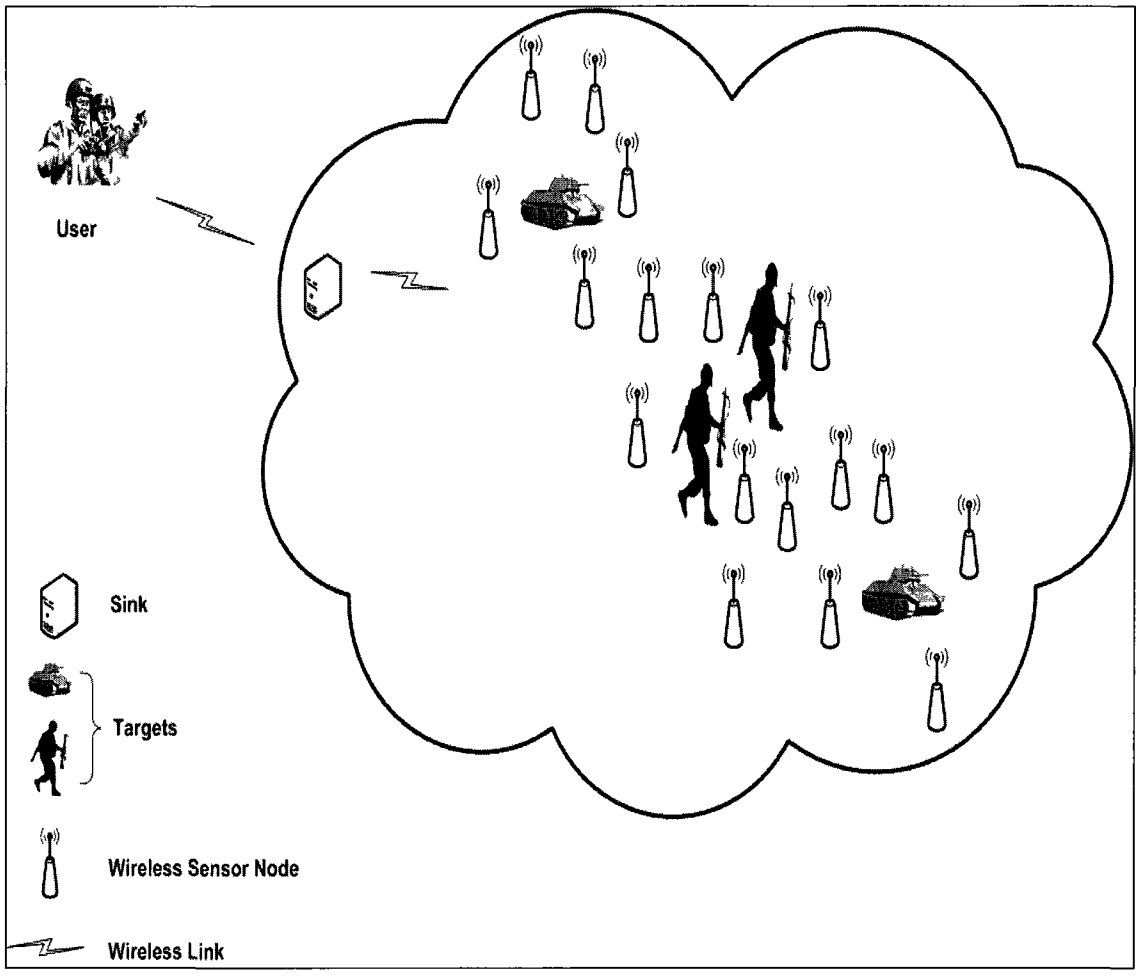

Figure 2.2: An example of WSN application in POCW topology: The wireless sensor nodes are tasked to monitor the battlefield targets (Troupes and Tanks). The wireless sensor nodes send sensed data to the the sink using a wireless channel. Then, the sink relays the data to command and control station for decision making.

These sponsor sets would be defined in such a way that each one is capable of covering all the targets within the field of interest. A global schedule is then defined to identify the period in which each sponsor set is to be active. Defining the sponsor sets is not an easy task but several solutions have been proposed for this problem by introducing different heuristic techniques (see [23]) and this task is not a part of the focus of the thesis.

2.1.2 Observations and Events

The purpose of a WSN is to provide sensing capabilities for detailed monitoring of a phenomena. We will refer to the low-level readings from the sensor nodes as *Observations.*

As mentioned before, the continuous operation of sensor nodes generates an overwhelming volume of observations that need to be either transmitted to the network user or stored somewhere in the WSN (locally or remotely). In terms of energy cost, the transmission of these data to the sink node (or user), whether for storage or for back up purposes is prohibitive. These operations would quickly drain the power reserve of any sensor nodes and by extension shorten the network lifespan. However, this likelihood can be avoided or at least reduced by observing that network's users are generally not interested in all available raw observations. Network's users are rather interested in a specific subset of observations that are meaningful when processed. We, then, define an event to be a subset of the observation set which meet a well-defined user criteria [24].

Events within the covered area can thus be detected by processing the observations. Events can be defined not only in terms of observations but also in terms of other events. These events may not be in a strict hierarchy. However, some events may be of lower level than others depending on the application and the context, and then the lower level events may be used to define other higher level events [19].

For example, data readings from a POCW used in battlefield surveillance, like the one shown in figure 2.2, would be observations while analysis of these observations to detect enemy sighting on the battlefield might lead to an **alert** event. Furthermore,

continued surveillance could yield the enemy's activity records which when processed might show a particular increase in intensity and frequency of alert events which would signal an **imminent attack** event.

2.1.3 Scalability and Robustness

2.1.3.1 Introduction

The scale of WSN is wide ranging, and it may be a factor of sensor nodes' density and the size of the covered field of interest. Density varies from low, with few sensor nodes deployed, to high, with millions of sensor nodes deployed. An example of high density is the smart dust [25]. Covered field size may be from a room size, in a house, to a very wide physical region [1] like a continent. In any case, energy constraints are severe since sensor nodes operate on battery power, independent of the number of sensor nodes involved. Once the network is deployed the number of sensor nodes may vary a lot but the network is expected to remain functional despite this variation. It then becomes imperative to ensure robustness of the network throughout its useful lifetime. Therefore, the network scalability and robustness must be preserved at the same time in a WSN even though the network is distributed by nature.

2.1.3.2 Design Criteria for Scalability and Robustness

The following criteria have been suggested when designing scalable and robust WSN [19].

1. Scaling in sensor node: average communication energy cost should not significantly increase with the increasing number of sensor nodes. Nor should any node become a concentration point of communication.

2. Persistence: any event record k should be available to WSN as long as it has not been out of date despite sensor node failures and changes in the sensor network topology.

3. Consistency: a query about event k must be routed correctly to a sensor node v where information about k is currently stored; if v is later on replaced by w then query should reflect this change to maintain persistence; after a node failure, queries and stored data must be routed to the replacing sensor node consistently.

4. Database scalability: if data are backed up, data should not be backed up at one node.

5. Topological generality: topology should not be a limiting factor.

2.1.4 Summary

WSN allow remote but accurate monitoring of physical environment. WSN are often deployed in inhospitable and nearly inaccessible environments. These environmental and accessibility constraints make resource management an essential requirement. This introduction reviewed WSN and looked at some of the issues that must be dealt with for successful operation. The next section introduces the process of knowledge discovery in database.

2.2 Knowledge Discovery In Database

The value of data is no longer in how much of it you have The value is in how quickly and how effectively can the data be reduced, explored, manipulated and managed

This section provides a brief introduction of the knowledge discovery process. The section is organized as follows. Section 2.2.1 provides a glimpse of the reasoning behind knowledge discovery process. Section 2.2.2 defines the knowledge discovery process. Section 2.2.3 describes the phases of knowledge discovery process. Finally, Section 2.2.4 summarizes the section.

2.2.1 Introduction

According to Lyman (2003), if the amount of information in the world doubles every 20 months, then the size and the number of database to store that information probably increases at an even higher rate. In addition, information flows at such a high rate that it cannot be viewed or analyzed by humans for the foreseeable future [26]. Consequently, the gap between information generation and information understanding would grow quickly without the use of automated analysis tools [27]. Existing tools consisting of computers using the same techniques as humans, like simple statistical techniques for data analysis, take time to yield results. Therefore, there is a need for fast and intelligent data analysis techniques. In the knowledge economy of today, the ability to create new knowledge fast represents a competitive advantage [28] thus timeliness is key in information analysis. This implies that information analysis should be carried out by very powerful and intelligent computers.

The increasing number of deployed WSN, and by extension the number of deployed sensor nodes, result in huge amount of sensed data that need to be processed. Evidently, raw data must be analyzed to make data collection worthwhile. In addition,

the size and the complexity of data transmitted by sensor nodes grow exponentially as new sensor nodes track their environments with better precision. Analysis of this growing volume of data is a challenging task because not only the environment surrounding the sensor nodes evolves over time but also because of the complexity and the size of gathered data [29]. From this reality comes the necessity to use computer tools and develop new approaches that can handle large amount of sensed data and, hence, to help us understand monitored phenomena.

The quest for knowledge has created a knowledge management practices to identify, create, represent, distribute and enable adoption of knowledge. The aim of Knowledge management in WSN is to discover, explore, and manage knowledge embedded within sensed data. Knowledge management consists of [10]: *Knowledge Repository, Knowledge Sharing, and Knowledge Discovery.*

Knowledge repository is the creation, storage and management of sensed data.

Knowledge sharing involves the exchange of sensed data.

Knowledge discovery includes the analysis or the mining of data for knowledge.

The use of Knowledge Discovery in Databases in WSN is a knowledge management effort to deal with data volume and data complexity issues.

2.2.2 KDD Definition

According to [30], Knowledge Discovery in Databases, hereafter abbreviated KDD, describes the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. This means that the ultimate objective

of the KDD process is to extract new knowledge from data. The process may be iteratively refined until new knowledge is discovered.

The term *knowledge* in KDD emphasizes the notion of *new*, *nontrivial*, and *useful* information (patterns) from data [30]. The process should result in *new* information (patterns) at least to the system and more importantly to the user. Discovered information should be *nontrivial* in the sense that information (patterns) cannot be obtained by simple statistical techniques. This implies that a KDD system has some degree of autonomy for data processing and result evaluation. Finally, discovered knowledge must be *useful* as to benefit the user. This means the information (patterns) must be relevant and understandable in light of the WSN goals.

In practice, the above three notions are named *interestingness* and *certainty* [31]. Interestingness provides a clearer and overall measure of information (patterns) value. It is a function of validity, novelty, usefulness, and simplicity. It can be easily defined explicitly or implicitly by the user. Hence, a new information (patterns) may become new knowledge if it exceeds some interestingness threshold. Depending on applications, the user is able to fully define what constitute new knowledge simply by choosing whatever functions and thresholds deemed appropriate [30]. Certainty, on the other hand, represents the amount of faith to accord to the newly discovered knowledge. It depends on data integrity and on the size of data samples used in the KDD process [28]. Uncertain information (patterns) cannot be proved and is not considered to be knowledge.

Finally, we have to keep in mind that KDD can only be a replacement option to traditional analysis technique if its results are timely. Timeliness implies that KDD processes should be efficient in terms of computational complexity [28]. In other

words, the process running time must be predictable.

2.2.3 KDD Phases

According to [30] the KDD process has three main phases:

1. Data Pre-Processing
2. Data Processing
3. Data Post-Processing

Each phase comprises several steps that can be completed iteratively as illustrated in figure 2.3.

Data pre-processing is concerned with the raw sensor data. Its main goals are understanding the objective of the WSN, the nature of sensed data, specifying sensor data requirements, and finally sensor data preparation. Data processing, also name data mining or modeling, tries to extract hidden patterns from sensor data. Data post-processing is the evaluation, interpretation, understanding, and refinement of data mining result [13, 10].

2.2.3.1 Pre-Processing

Pre-processing is structured as follow [10]: define objective, find sensor data characteristics, sensor data preparation.

First, the objective of KDD process must be understood. Once the objective of KDD process is known, sensor data characteristics and specifications are formulated to identify the data source, type or format, quality, etc. Domain knowledge plays

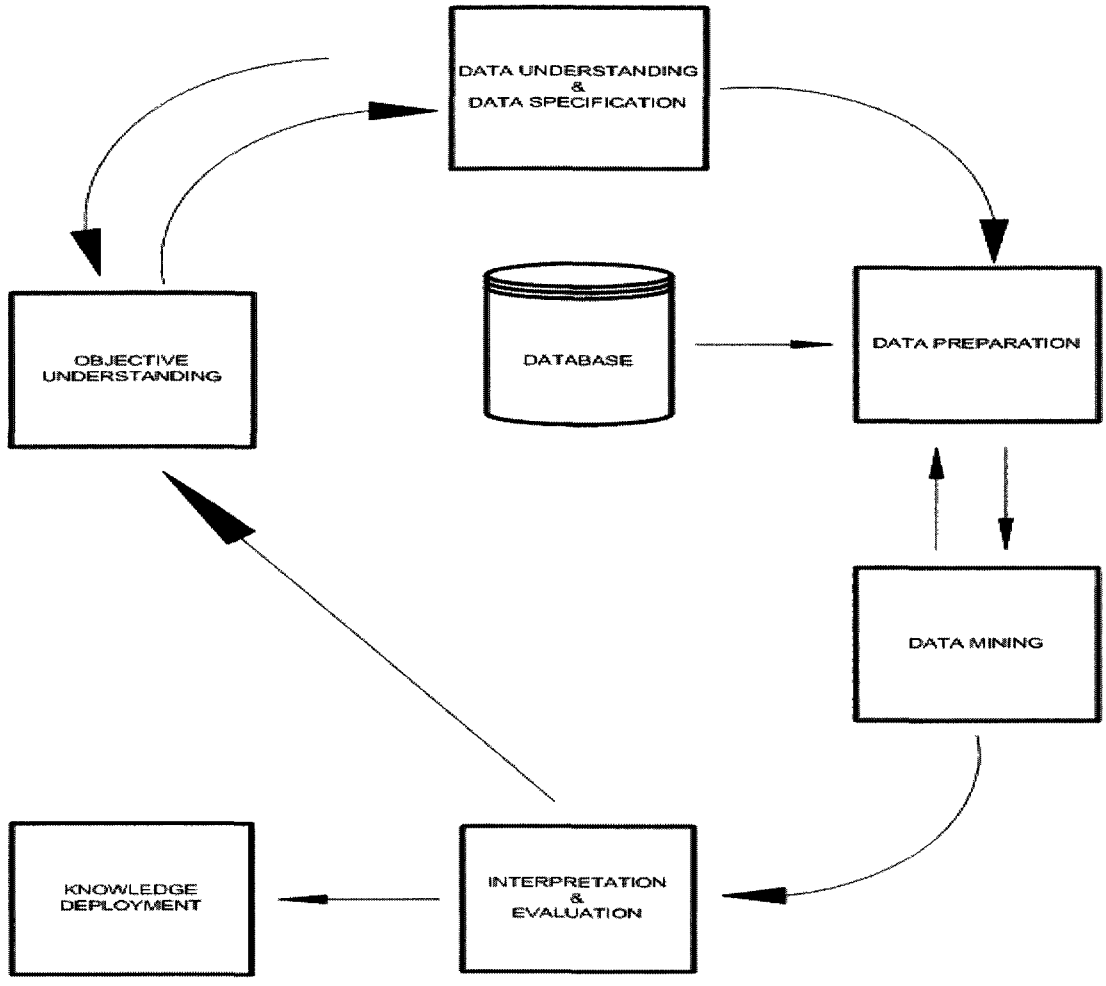

Figure 2.3: KDD Process Phases: Pre-Processing (Objective Identification, Data Transformation, and Data Preparation), Processing (Extract Interesting Patterns from Data), and Post Processing (Interprets and Evaluates Extracted Patterns). Note that the Process is iterative. Image from CRISP-DM 1.0 Step-by-step data mining guide (2000)

a critical role at this stage of pre-processing. The focus is mainly on the data type (numeric, binary, continuous, etc) of the collected data because mixed data type may slow down the data mining phase. In addition, the data mining phase may mishandle mixed data types by treating all data values indiscriminately [13, 11, 10]. Data quality deals with elements of accuracy, relevance, and completeness [13, 11, 10]. The WSN user should ensure that the quality of collected data will support the KDD objectives.

The last element of the pre-processing is *sensor data preparation.*

Sensor Data Preparation is an important part of the KDD process carried out right before data mining [13, 10]. Once the data quality and characteristics have been assessed, significant effort is put toward the preparation of data before its analysis. Data preparation tries to reveal the embedded content of raw data through manipulation and transformation in an attempt to facilitate knowledge discovery [10].

During data preparation, sensor data are also cleaned for example by removing outliers, duplicate, and missing sensor data [32]. In order to reduce the processing effort, data preparation tries to significantly reduce the overall size of sensed data to be mined to a manageable but representative sample. The reduction depends on the KDD objective. Obviously, selecting the right data preparation criteria is critical for the success of the following phases [32].

To highlight the importance of clean and relevant data, figure 2.4 illustrates the KDD process steps and the relative effort typically associated with each step. As shown, 60% of the overall effort is spent on data preparation while the actual mining step represents only about 10% of the overall effort. Thus, data preparation is one of the most important steps but also the most difficult and time consuming in the entire KDD process [33].

Clearly, data preparation effort generate non negligible overhead. Consequently, the associated costs and benefits must be estimated to justify the investment in sensor data preparation. In this thesis, the cost and benefits respectively refer to the energy consumed during data preparation and the discovered knowledge. Therefore, We will propose three algorithms to minimize the energy cost of performing data preparation

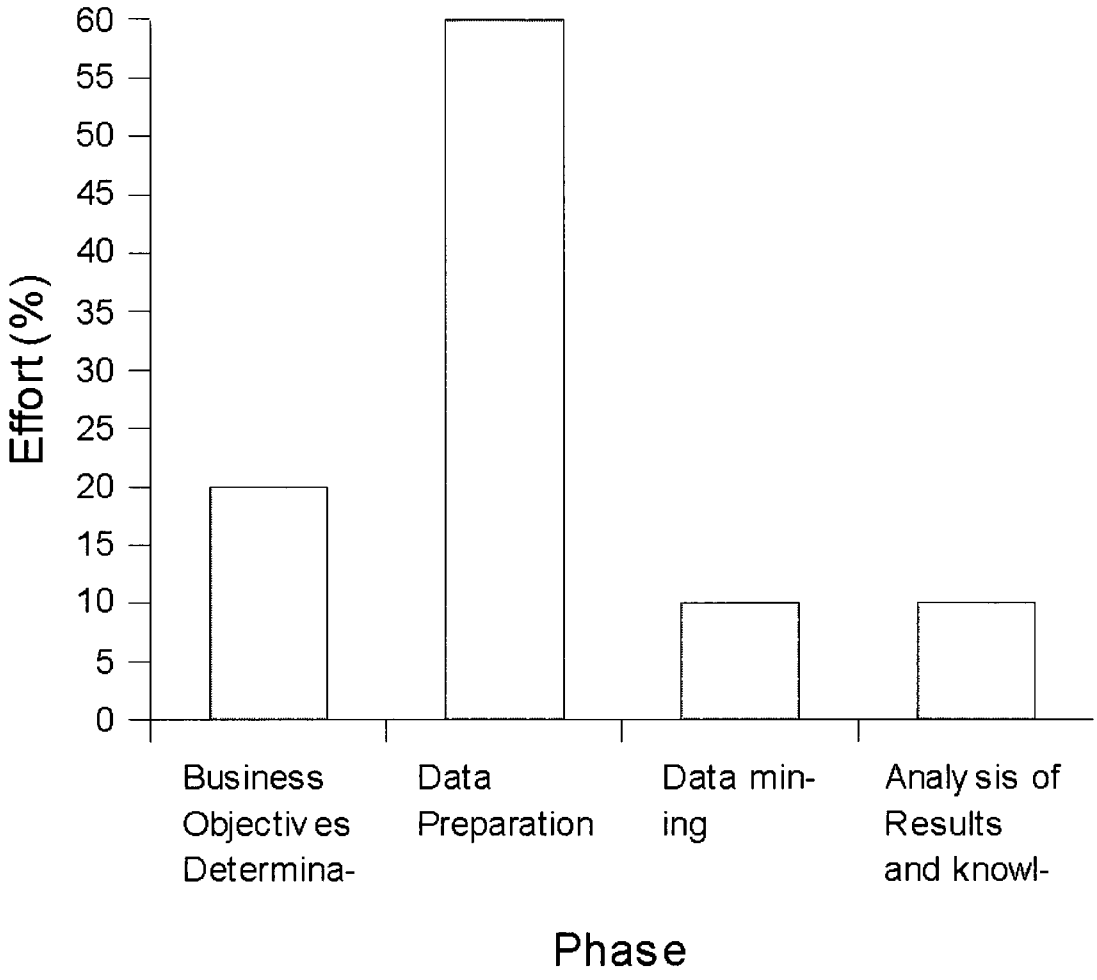

Figure 2.4: Effort Required for Each Data Mining Process Step [12], 20% of the overall effort is spent defining KDD objectives, 60% of the overall effort is spent on data preparation alone, 10% of the overall effort is spent on data mining, the last 10% of the overall effort is spent on analyzing and refining mined results. Data Preparation step requires the most effort i.e. it is the most time consuming and difficult step of the entire KDD process

for POCW.

2.2.3.2 Data Processing

Why Data Processing? Because of two reasons: The first reason is the increasing rate of growth of new data which can no longer be dealt with in timely manner using

traditional analysis techniques even with the help of computers. The second reason, and the most important, is the possibility of discovering new information (patterns) without previously formulating a hypothesis [12]. Data mining techniques provide a mean to visualize or understand large volume of data with multiple dimensions, and the user can specify queries in more abstract ways than currently possible [27].

Data can grow in dimensions by increasing the number of fields (attributes) and the number of cases. The traditional approach of dealing with high-dimensional data is the *divide and conquer* method of reducing the data into blocks of lower and simpler dimension that can be easily analyzed. However, this solution does not scale since the number of possible combinations for dimensionality reduction grows exponentially with the number of dimension, thus rendering the optimization of dimensionality reduction or the grasp and view of the complete solution impossible. A better way to overcome these hurdles uses data mining to perform the appropriate dimensionality reductions [27].

Data Mining tries to extract hidden patterns from sensor data [13]. It is concerned with the algorithmic means by which patterns are extracted from sensor data and enumerated [28].

Formally we can say that *given a set of data (facts) DB, a language L, and a measure of certainty C, we define a pattern as a statement S in L that describes relationships among a subset D of DB with a certainty C, such that S is simpler than listing all data in D.* If a pattern is *interesting* (i.e. meets the interestingness threshold) and *certain enough* (i.e. meets the certainty threshold), it becomes *discovered knowledge* [28].

In this thesis, sensor data are sets of observations about some target's activities while patterns are expressions representing a restrained description of a subset of the sensor data [27], so extracting a pattern means finding structures in sensor data. In general terms, this is the same as describing sensor data in some high-level language. The space of possible patterns may be infinite [27]. This is why constraints (interestingness, certainty, etc.) are used to limit data mining algorithms to patterns from particular subspaces.

Data mining commonly involves four classes of task [30]: *Classification, Clustering, Regression* and *Association rules. Classification* tries to arrange the data into predefined groups. *Clustering* is considered a general case of classification. It simply groups similar items together without using predefined groups. *Regression* attempts to find a function which models the data with the least error. Finally, *Association Rules* searches for correlation between data variables or attributes. Association rules will be defined in more details later on.

2.2.3.3 Data Post-Processing

Once data mining is completed, the extracted patterns are evaluated and interpreted because not all extracted patterns are necessarily valid. Therefore, patterns must be tested on data sample different from the one used by data mining algorithms. If the resulting patterns differ significantly from the data mining ones, then both the pre-processing and the data mining phases must be refined and repeated. If both set of patterns are similar, the final step is to interpret the patterns to turn them into knowledge [30].

2.2.4 Summary

As new WSN are deployed, they gather and store more data in their databases. as a result, the questions of data exploration and analysis becomes primordial. Knowledge Discovery in databases (KDD) is a process which seeks to extract useful knowledge from vast Data. KDD offers the capacity to automate complex search and data analysis tasks. Data mining is the main phase in the knowledge discovery process. It consists of the extraction of pertinent patterns hidden in data. The extracted knowledge is then used in the verification of hypothesis or the prediction and explanation of knowledge. However, KDD process cannot succeed without a good data preparation because data preparation enhances the data and facilitate data mining. We also need to know that several iterations of the KDD phases may be necessary to discover new knowledge. The next section will focus on the use of KDD techniques in WSN to identify behavioral patterns and improve their performance.

2.3 Review Of The State Of The Art

2.3.1 Introduction

This section presents the major and relevant ideas in the state of the art research in the application of KDD process in WSN data. The section is organized as follows. Section 2.3.2 presents the state of the art association rules in WSN. Section 2.3.3 presents the the state of the art data preparation mechanisms in WSN. Section 2.3.4 summarizes the section.

2.3.2 State Of Art in Association rules

This section present the major ideas in the state of the art research for discovering association rules in WSN. These ideas are grouped according to the objectives of association rules. Recall that association rules show patterns in sensor data which may describe one of two things:

1. *The Surrounding Environment*: patterns extracted from the sensor nodes readings, i.e. measurement of monitored phenomena.

2. *The Sensor Behavior*: patterns extracted from meta-data describing sensor nodes behavior, i.e. about sensor nodes themselves.

2.3.2.1 Patterns between sensor nodes environment

Several works have adapted Association Rules to WSN depending on their intended applications. Mainly, these works have exploited the data values of the sensor nodes. In other words, the sensor nodes readings have been the main objects of the rules. For example, [34] considered the issue of mining association rules in data streams generated from sensor nodes in a WSN. In [35], the authors considered the distributed nature of WSN and proposed an in-network data mining technique that discovers frequent spatial and temporal patterns of events. In [36], the authors proposed an association rule mining framework that tolerates missed readings due to message loss or data corruption during routing from sensor nodes to the Sink. In [18], the authors proposed a light weight mining algorithm for energy conservation. the same objective is achieved in [37] by using in network management. This serves as an illustration

of the wide application of association rules when dealing with readings from sensor nodes in a WSN. Patterns between monitored environment are not our main interest for this thesis, the interest lies in patterns between sensor nodes and targets.

2.3.2.2 Patterns between sensor nodes behavior

Patterns between sensor nodes behavior capture the temporal relations between sensor nodes based on their activities without considering their actual readings.

In [14], the authors proposed the Sensor Association Rules as a way to capture the temporal correlation between sensor nodes in a WSN. Note that sensor association rules differs from the association rules described in subsection 2.3.2.1 because, in regard to sensor association rules, the sensor nodes themselves are the main objects of the extracted rules not their readings. In addition, sensor association rules capture the temporal relations between sensor nodes activities without considering the order in which the events were detected. In [14], the authors also proposes a data structure, called the Positional Lexicographic Tree (PLT), to compress sensor data and for efficient mining of the sensor association patterns.

Also of interest is a modification to the sensor association rules named the Coverage-based association rules, these later rules capture the correlation between a set of locations monitored by a WSN network. The difference with respect to Sensor Association Rules is that Coverage-based Rules are patterns between locations in the WSN's field of interest rather than patterns between sensor nodes. The underlying assumptions being: first, the partitioning of the field covered by the WSN into a set of locations, each covered by k-sensor nodes, where $k >= 1$; and second, the election of a location manager responsible for collecting and reporting all of the behavioral

data sensed in each location [8].

The authors of [38] proposed another version of the sensor association rules capable of capturing chronological patterns i.e. temporal correlation between detected events, more specifically the order of event detection.

2.3.3 State Of Art in Data Preparation

Data preparation is an important part of KDD carried out right before data mining [13]. Once the data quality and characteristics have been assessed, significant effort is put toward preparation of data before analysis. Data preparation tries to reveal the embedded content of raw data through manipulation and transformation in an attempt to facilitate knowledge discovery [10].

In order to extract any type of Sensor Association Rules, sensor nodes must record and provide their own behavioral data (meta-data) to the sink or to any other entity responsible for mining the rules. This step is the data preparation of the recorded meta-data, in other words, the recorded sensor nodes activities over time. The meta-data describes sensor behaviors and are different from sensor nodes readings of the surrounding environment. By providing this information, the sensor nodes inform the Sink about the time slots in which events were detected.

Impacts of data preparation on knowledge discovered in terms of accuracy, completeness, relevancy, and timeliness were discussed in [13, 11]. However, the aim of these articles was the evaluation of the impact of data preparation on the discovered knowledge itself; and, none of these algorithms was designed to collect WSN data for mining purpose. They did not consider the redundancy in behavioral data from

sensor node to sensor node. These factors force the need for special data collecting techniques for mining purpose. with special focus on how to efficiently extract the behavioral data from resource limited sensor nodes. Behavioral data could then be collected using existing routing protocols and data gathering algorithms.

Different mechanisms for data preparation were proposed to take advantage of the stored sensor behavioral data. The proposed solution required storage capacity on each sensor node to allow local profiling and processing of sensor nodes' activities before any transmission to the sink node. In [14, 8], this processing involved compression techniques that reduced the amount of data transmitted by each sensor node with the stated aim of reducing the energy consumption in WSN. In the case of generating Sensor Association Rules [3, 14, 15], the authors proposed either a (i) *Direct Reporting mechanism* in which behavioral data of each sensor node are transferred to the Sink without any processing by the sensor node, or a (ii) *Distributed Extraction mechanism* in which more computational load is carried by the sensor nodes. In [15], the authors proposed an in-Network Reduction mechanism that minimizes the number of messages needed to encapsulate the behavioral data by eliminating redundancy in those behavioral data. [38] presented a data preparation technique for generating Chronological Patterns. The authors used a tree structure, named the Chronological Tree structure model, to compress stored sensor nodes behavioral data. The chronological Tree reduces the dimension of the combination and the complexity of the sequence to be checked during data mining [38].

2.3.4 Summary

To summarize, inclusion of energy awareness in data preparation mechanisms could help extend the network lifetime by reducing the energy consumption associated with data preparation. The objective of this thesis is to find energy aware data preparation mechanisms to be used during knowledge discovery in the context of POCW.

We can see that although the importance of data preparation in the KDD process is recognized in WSN [10, 14, 8, 39], there is no evaluation of energy consumption associated to data preparation stage in the context of POCW. In the next chapter we will consider three different transmission approaches. We will begin by a description of these approaches followed by a performance evaluation from simulations.

Chapter 3

TARGET-BASED ASSOCIATION RULES FOR POCW

3.1 Introduction

In this chapter, we present a special type of sensor association rules, the Target-based association rules for Point-of-Coverage WSN. These new association rules are derived from the Sensor Association Rules proposed in [14]. These rules may be used for WSN management in order to improve the Quality of Services of WSN. As explained in section 2.2.3.2, sensor association rules result from data mining.

We will also consider the data preparation step in the KDD Process. This step features the extraction mechanisms required to collect data from sensor nodes. In our context, data used in this process refers to sensor node meta-data, and differs from the actual (raw) readings of the sensor nodes. We propose three different mechanisms to be used in the data preparation step. Our objective is to minimize the energy cost

associated to the data preparation by eliminating redundancy in transmitted data, this effectively reduces energy cost of communication. The prepared data are then used by the sink to generate Target-based Association Rules.

This chapter is organized as follows. Section 3.1 provides a general introduction to this chapter. Section 3.2 outlines some of the assumptions made in the solution development. Section 3.3 provides a brief definition for the Sensor Association Rules and provides a formal definition of Target-based Association Rules. Section 3.4 defines the proposed solutions. Section 3.5 describes the proposed low energy data preparation mechanisms used to prepare the data needed for generating target-based association rules. Section 3.6 outlines some of the limitations of the proposed solutions. Section 3.7 summarizes the chapter.

3.2 Assumptions

Before we go any further, a number of assumptions must be made.

First, we assume throughout this chapter that the words network and WSN means Point-of-Coverage wireless sensor network and all sensor nodes within this network know their geographic location. This can be achieved by using one of the several localization techniques [40, 41]. This assumption is not critical for our proposed algorithms; however, it may be more useful for the network user to know the source location of sensed data.

Second, we assume that each sensor node is able to establish and maintain a communication link with the sink node until its energy reserves no longer permit that. In field application, sensor nodes in the vicinity of the sink would use a direct

link, while those farther away would route their message to the sink node through some access points; then, to complete the path to the sink would require sensor nodes to route their messages to an access point first then from there to the sink node. This assumption is required to compare the proposed data preparation mechanisms and to permit relative comparison with other existing data handling mechanisms.

Third, we assume that energy is a scarce commodity for all sensor nodes [1], and that data handling algorithms should seek to minimize energy cost of communication operations in order to extend the overall network lifetime.

Fourth, we assume each sensor node can measure its energy reserve with accuracy, and can predict its remaining lifetime based on past activity records [42].

Fifth, we assume the existence of a reliable wireless communication link between the sensor nodes and the sink. While the mapping between communication and energy consumption is complicated, we will use the first order radio model described in [5].

Sixth, the database used by the mining algorithm is made from the data sent by the sensor nodes, it is important to note that this database may be centralized in one place (at sink node for example) or distributed (at the sensor node level).

For performance evaluation purpose, two metrics are used: the *Average Energy* usage and *Total Number of Message*. *Average Energy*, computed per unit of time, is the total amount of energy consumed in the networks in one unit of time divided by the number of sensor nodes scheduled in that same unit of time. *Total Number of Message* is the total number of messages exchanged in the WSN.

While we treat all sensor nodes as thought they have the same initial capabilities prior to deployment, the reality may be otherwise. Here, we neglect the likely evolution of many WSN into a tiered architecture. Some nodes may have very limited

resources while others have much more significant resources (extra battery and data storage). For example, newly deployed sensor nodes have more battery capacity and better communication equipments due to the evolution of embedded technologies over time.

3.3 Sensor Association Rules

An association rule is of the form If S Then P, where S is the condition of the rule and P is the prediction. In the context of WSN, S may be a set of sensor nodes or targets and P may be a single sensor node or target. Sensor Association Rules attempt to capture the temporal correlation between sensor nodes of a WSN [14]. In general terms, it is the set of sensor nodes that detect events in common time intervals. Sensor association rules are generated during the data mining phase during the KDD Process described in 2.2.3.2. These rules may be used for WSN management in order to improve the Quality of Services of WSN.

Recall that, in this work, we want to define behavioral patterns between the different targets under monitoring. A direct solution would apply the same methodology used in [3] where temporal relations are defined between the sensor nodes in the network. However, this solution is not optimal because:

1. Sensor Association Rules captures relations between all the sensor nodes of the WSN, but we are instead interested in capturing patterns between a sets of targets in a POCW.

2. Generating Sensor Association Rules requires the sensor nodes to be always ac-

tive but this would limit our ability to use energy-aware data gathering mechanism.

3. A redundant data would be delivered to the Sink.

These three concerns point to the need of a more appropriate framework for capturing relations among targets in the context of POCW. In the next section, we propose the Target-based Association Rules as the main framework to generate behavioral patterns among a set of targets in POCW.

3.4 Solution Definition

In this section, we present the Target-based association rules for POCW. This form of association rules attempts to capture the temporal correlation between targets present in the field of interest of a POCW, or simply the set of target that are active in common time intervals. Such rules can be used for surveillance and target tracking applications.

3.4.1 Target-based Association Rules Formal Definition

Target-based association rules are based on the common intervals of events occurrences at the targets. In order to generate association rules between a set of targets, we have to collect behavioral data that describes the activity of these targets over time. We now present a formal definition for all the main concepts needed to generate Target-based association rules.

Let T = $\{T_1, T_2, ...T_n\}$ be a set of targets in a particular POCW field of interest.

Let S = $\{s_1, s_2....s_m\}$ be the set of sensor nodes deployed within this field. We assume that each sensor node is responsible for covering a set of targets in the network. We use the same parameters used in [3] to define Sensor Association Rules. Those parameters were the slot size (λ), the historical period (T_his), and the minimum support.

To be able to generate these association rules, time is assumed to be divided into a set of time slots of size λ each. The historical period (T_his) is the period of time to profile the targets' activities in other word how long the targets' behaviors should be recorded, and it is determined by the application or the user. Each sensor node is responsible for profiling the activity of all targets within its coverage area. To record these profile for later analysis, sensor nodes maintain a set of storage buffers, one for each covered target. These buffers, referred to as *behavioral buffers*, are used to encode the time periods (more specifically the time slot numbers) in which activities were detected at each target. Each buffer contains an entry for each time slot within the given historical period, therefore, each buffer will have a size equal to (T_his/λ).

Table 3.1: Target T_j Behavioral Buffer

1	0	0	1	0	1

During sensing operation, once activity is detected from a particular target, the sensor node takes the buffer reserved for that active target and set the buffer entry corresponding to the current time slot. For example figure 3.1 shows the content of a buffer corresponding to an arbitrary target T_j profiled by sensor S_i (of length 6). This buffer shows that activities have been detected at target T_j at time slots 1, 4,

and 6 because the first, fourth and sixth entries of that buffer have been set.

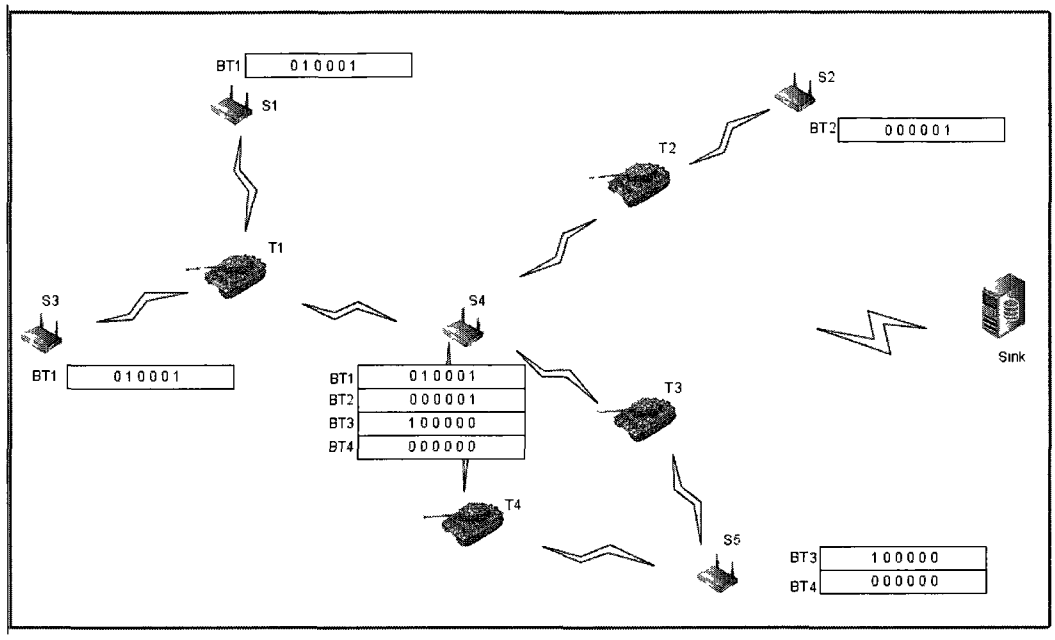

Figure 3.1: POCW Network with Sensor Nodes Showing Recorded Target Profiles, Point-of-Coverage Wireless Sensor Network, T_n : Target n, S_m : Sensor Node m, Sink : Sink Node. Each Sensor Node Maintains a Buffer for All Covered Targets

Figure 3.1 illustrates the concept of target profiling. We have a set of targets $\{T_1,\ T_2,\ T_3,\ T_4\}$ and a set of sensor nodes $\{S_1,\ S_2,\ S_3,\ S_4\}$ deployed to monitor the targets. We can see that sensor node S_1 covers target $\{T_1\}$ only, while S_4 cover targets $\{T_1,\ T_2,\ T_3,\ T_4\}$, and so on. We can also see, from the same figure, the content of behavioral buffer corresponding to an arbitrary target T_j profiled by sensor S_i. For example, the S_1 buffer shows that activities have been detected at target T_1 during the time slots 1 and 6 because the first and sixth entries of that buffer have been set. This figure also illustrates the redundancy of the data that may exists for each target in the field.

Definition 1. *Let T_i be a target in a particular POCW. $AS(T_i) = \{t_1, t_2, ..., t_m\}$, such that $m <= (T\_his/\lambda)$ and $B_{T_i}(t_j) = 1$, for all $1 <= j <= m$, is then defined as the Activity Set of Target T_i.*

Definition 2. *A Target Behavioral Database (TD) is defined as the set of targets, covered by a particular WSN, along with their Activity Sets.*

The following table shows an example of a Target Behavioral Database which is the database used by the mining algorithms. It is made from the data sent by the sensor nodes (i.e target buffer contents), it is important to note that this database may be centralized in one place (at sink node for example) or distributed (at the sensor node level).

Table 3.2: Target Behavioral Database for the POCW shown in Figure 3.1

Target	Activity Set
T_1	$\{2,6\}$
T_2	$\{6\}$
T_3	$\{1\}$
T_4	$\{0\}$

Definition 3. *$P = \{T_1, T_2, ..., T_k\}$, such that $P \subseteq T$, defines a pattern of targets.*

Definition 4. *The support of pattern $P = \{T_1, T_2, . ., T_m\}$ in the Target Behavioral Database is defined by the cardinality of the set that is produced by intersecting all the activity sets of the targets in this pattern.*

$$Support(P) = |\bigcap_{\forall l_j \in P} AS(T_j)|$$

Definition 5. *A pattern P is said to be frequent if its support is greater than or equal to a given minimum support.*

Definition 6. *Target-based association rule is defined as the implication $P' \Rightarrow P''$ where $P' \subset T$, $P'' \subset T$, and $P' \cap P'' = \phi$.*

Definition 7. *The support of the Target-based association rule ($P' \Rightarrow P''$) is defined as the support of the pattern ($P' \cup P''$) in the Target Behavioral Database, while the confidence of the rule is defined by:*

$$\textbf{\textit{Conf}}(P' \Rightarrow P'') = \textbf{\textit{Support}}(P' \cup P'') \; / \; \textbf{\textit{Support}}(P').$$

Knowledge Discovery in a target behavioral database is the process of generating all the Target-based association rules that meet a pre-defined minimum support and confidence percentage. In the next section, we will focus on data preparation mechanisms to be used in the process of creating Target behavioral database.

3.5 Target-Based Association Rules Data Preparation

Assuming the network has been deployed and all its elements initialized without a problem. Then, the sink knows how many sensor nodes are working as expected, their position and the position of targets in the field of interest.

First, the sink must divide sensor nodes into two *Sponsor Sets*. Recall that sponsor sets are disjoint groups of sensor nodes that completely and reliably cover all of the targets that must be monitored by the POCW. There is no restriction on the number of targets that any sensor node is allowed to cover and not all sensor nodes must be used provided all the targets are covered preferably without coverage redundancy. The method used for sensor nodes grouping is the one defined in [23]. Second, the sink schedules these sponsor sets for target monitoring, each sponsor set is scheduled for only half of the total historical period. The first sponsor set is scheduled for the first half of the historical period and the second sponsor set is scheduled for the second half.

In order to generate any Target-based Association Rules, the sink maintains target behavioral database which encode the activity set of all targets. The data in these database may be prepared either at the sink level or at the sensor node level. If the data is prepared at the sensor node level, the last step of data preparation would then involve the actual transmission of the prepared data to the sink where data mining is performed. Inclusion of energy conservation in data preparation mechanisms would help extend the network lifetime simply by reducing the energy consumption associated with data preparation. Toward this end, we propose three data preparation mechanisms for mining Target-Based association rules in POCW. We refer to these mechanisms by All-Nodes, Schedule-Buffer, and Fused-Schedule-Buffer data preparation mechanisms. All three mechanisms take into account the sensor nodes scheduling and exploit the nature of POCW to reduce energy cost of data preparation.

3.5.1 All-Nodes based Data Preparation Mechanism

This mechanism is similar to the Direct Reporting mechanism proposed in [3] for collecting the data needed to generate Sensor Association Rules. Recall that POCW uses sponsor set to schedule its sensor nodes, and that only one sponsor set is scheduled for active duty at any one time. The scheduled sponsor set monitors the activities of all the targets during its scheduled period of activity.

The All-Nodes based Data Preparation mechanism requires each sensor node in the scheduled sponsor set to be always active. Sensor nodes are assumed to have no storage buffers to record sensed data; therefore, no data pre-processing is performed prior to reporting detected event to the sink. Each time an active sensor node detects activity from one of its covered targets, it immediately sends to the sink a notification message carrying the time stamp of the observation time along with the target identification. The Sink accumulates all these messages over the historical period and uses them to construct the Activity Set of all the targets. Once the full activity sets for all the target are obtained, the sink creates a link to that Activity Set in the target behavioral database. Then data preparation is performed by the sink from the target behavioral database. Algorithm 3.5.1 shows a formal description of the All-Nodes data preparation mechanism.

Although the All-Nodes data preparation mechanism sounds simple and no overhead is put on sensor nodes, it is a costly solution in terms of energy consumption considering the energy constraints of sensor nodes. Generally, all transmitted observations will not be used in the formulation of the Target-based association rules because some of these targets activities may not be interesting. Recall that the ob-

jective of data preparation is to reduce the size of data to be mined to a manageable sample, and that the application wants to capture *highly interesting* association rules only. Clearly, this is not the case because every observation data is transmitted to the sink. In fact, no data preparation is done in the sense of reducing the transmitted data for efficiency or energy conservation. It is the sink node that has to perform the actual data preparation before data mining.

Algorithm 3.5.1 All-Nodes Data Preparation

Initialize Database;
Upload data parameters;
while Network is Active **do**
  for Each Sensor Node **do**
    Target **Monitoring and Reporting**;
  end for
  for Each reported message M **do**
    Maintain Database
  end for
end while

Algorithm 3.5.1 is further explained below starting with algorithm 3.5.2 that maintains the target behavioral database (TD) as explained in the definition of Target-based association rules.

Next, algorithm 3.5.3 is used by the sink to upload the mining parameters into the sensor nodes in the network.

Then, algorithm 3.5.4 is used by sensor nodes to monitor and report their covered

Algorithm 3.5.2 Initialize Database

Sink Node:

$TD = TargetDatabase$;

$T = SetofCoveredTarget$;

for $EachT_i \in T$ **do**

B(T_i) = Behavioral Buffer of Target T_i;

end for

Algorithm 3.5.3 Upload data parameters

Sink Node:

Broadcast global schedule;

Broadcast mining parameters = { $T_{hist}, \lambda, min\_sup$ };

targets.

Algorithm 3.5.4 Target Monitoring and Reporting

Sensor Node:

T_{Sche} = Scheduled Time;

while Current Time $\leq$ Time + T_{Sche} **do**

if Event is detected **then**

Time Stamp = Current Time

M = { Sensor id, T_i, Time Stamp};

Send M to the Sink;

end if

end while

Finally, the algorithm 3.5.5 is used by the sink to update information in the target

behavioral database (TD).

Algorithm 3.5.5 Maintain Database

Sink Node:

Upon receiving all sensor nodes messages;

for Each message **do**

Extract T_i and Time Stamp;

Slot Number = $CurrentTime/\lambda$);

if $B(T_i)[slotNumber]$ not set **then**

Set $B(T_i)([slotNumber])$;

AS_{T_i} = The Activity Set of T_i;

Insert $\{AS_{T_i}$, TD $\}$;

end if

end for

Clearly, the All-Node based data preparation mechanism does not meet the energy reduction requirement. A better solution must consider and exploit the nature of the WSN at hand during data preparation in order to reduce the amount of observations transmitted to the sink. We already know that the WSN at hand is a POCW, and that all transmitted observations are not useful for data mining. In addition, the reported observations may be redundant if there is overlap in coverage area of sensor nodes in the same sponsor set.

The next two mechanisms will try to reduce data redundancy and energy consumption by coordinating data preparation activities among sensor nodes in the same sponsor set. Sensor nodes must store and pre-process sensed data to reduce the amount of

required transmission. The two mechanisms are referred to by Schedule-Buffer data preparation mechanism and Fused-Schedule-Buffer data preparation mechanism. The first mechanism considers sensor data collected during individual sensor node schedule time. Generally, the schedule time of a sensor node is less than the historical period, and allows only partial profiling of a target. The complete profile is only available when partial profiles of a target are integrated. The second mechanism fuses partial profile from individual sensor node so as to get a profile for the entire historical period at the sensor node level.

3.5.2 Schedule-Buffer based Data Preparation Mechanism

This mechanism redefines the All-Nodes data preparation mechanism by taking advantage of the grouping of sensor nodes into sponsor sets. We know that some targets may be covered by several sensor nodes in the field. With the All-Nodes data preparation mechanism, each sensor node in the scheduled sponsor set reported the activities of all their covered targets. Hence, targets covered by several sensor nodes were reported more than once. This resulted into redundant data at the sink.

The Schedule-Buffer based Data Preparation Mechanism eliminates this redundancy by making sure that each target behavioral profile is reported only by a single sensor node. This condition requires the sink to tell all sensor nodes of the scheduled sponsor set what subset of all of their covered targets to report. More importantly, the target profile is reported only if it meet the minimum support i.e. the interestingness threshold. Algorithm 3.5.6 gives a formal description for this mechanism.

Algorithm 3.5.6 is further explained below starting with algorithm 3.5.7 used by

Algorithm 3.5.6 Schedule-Buffer based Data Preparation

Upload data parameters;

while Network is Active **do**

for Each Sensor Node **do**

Target Monitoring;

Report Target Activity;

end for

for Each received message M **do**

Extract Activity Set AS_{T_i} ;

Insert { AS_{T_i}, TD };

end for

end while

the sink to upload the mining parameters into the sensor nodes in the network.

Algorithm 3.5.7 Upload Data Parameters

Sink Node:;

Broadcast global schedule;

Broadcast mining parameters { $T_{hist}, \lambda, min\_sup$ };

Each sensor nodes decides what to do according to algorithm 3.5.8.

Target monitoring consist of observing the target behavior and recording any observation using algorithm 3.5.9.

When reporting recorded target activities sensor follows algorithm 3.5.10.

Contrary to the All-Nodes data preparation mechanism, sensor nodes are assumed to have storage space to record targets' activities. The presence of storage allows

Algorithm 3.5.8 Target Monitoring

```
Sensor Node:;
Upon receiving mining parameters and schedule Time;
while Healthy do
  if Scheduled then
    Observe Covered Targets;
  end if
  if Not Scheduled then
    Go into Sleep Mode;
  end if
end while
```

limited pre-processing of collected data at the sensor node level before transmission to the sink. In addition, we assume that there is no direct collaboration between sensor nodes, that sensor nodes know the mining parameters, and that there exists a global schedule defined by the Sink. The global schedule defines the schedule time for each sensor node (i.e. the period in which the sensor node should be active). Also, the sink ensures that each target's profile is reported only once i.e. each target must be monitored by only one sensor node in the scheduled sponsor set. The sink select the targets to be covered by each sensor node from the set of all covered targets based on how close the target is relative to the sensor node. Sensor nodes only monitor and report behavioral profiles of its assigned targets. Once the targets to be monitored are known, each sensor node then maintains a set of buffers, one for each monitored target. Each buffer contains a buffer entry for each time slot in the given historical period.

Algorithm 3.5.9 Observe Covered Targets

T_{Sche} = Scheduled Time; Time = Current Time;

$T = SetofCoveredTarget$;

for *Each* $T_i \in T$ **do**

B(T_i) = Behavioral Buffer of Target T_i;

end for

while $CurrentTime \leq Time + T_{Sche}$ **do**

if event is detected **then**

Slot Number = $CurrentTime/\lambda$)

Let T_i be the target where event is detected and B(T_i) is the corresponding buffer;

if $B(T_i)[slotNumber]$ not set **then**

Set $B(T_i)([slotNumber])$;

else

Do nothing;

end if

end if

end while

Report Target Activity;

During the scheduled time, once activity is detected from a particular target, the sensor node retrieves the buffer reserved for that active target and set the entry corresponding to the current time slot. At the end of the scheduled time (which in this case corresponds to half of the historical period), each sensor node scans its buffers to count the number of entries that were set. If the count meets or exceeds the minimum support, those buffers are prepared and sent to the Sink.

Algorithm 3.5.10 Report Target Activity

Sensor Node:;

Require: **End** of scheduled time reached

for each target T_i within the sensor's coverage **do**

AS(T_i) = The set of time slots that have set bits in B(T_i);

if $cardinality of AS(T_i) \geq min\_sup$ **then**

M = { Sensor id, T_i, $AS(T_i)$ };

Send M to the Sink;

end if

end for

Without explicit collaboration among sensor nodes, the sink must task the sensor nodes in way that ensures that any complete target behavioral profile that meet the minimum support is always reported. The problem is that the complete target behavioral profile is obtained by merging partial behavioral profiles from two sensor nodes grouped in different sponsor sets. On one hand, one partial profile is reported from the sensor node scheduled in the first half of the historical period and the other partial profile from the sensor node scheduled for the second half of the historical period. There is no guarantee that if one sensor node reports a particular target's partial behavioral profile then the other sensor node does the same so that the sink has the complete behavioral profile of that particular target. Such case could occur if a target is particularly very active in one half of the historical period only and almost inactive in the other half. This means that, a sensor node scheduled when the target was not very active may not report its partial profile even thought the two partial profiles would meet the minimum support once combined.

The absence of collaboration makes the reporting less efficient. Each sensor nodes must report its partial profile independently without the knowledge of whether the complete profile meet the minimum support or not since they only have a half of the profile. But how to ensure that the needed partial profile is sent to the sink? Since the sensor node does not know before hand which target profile will meet the minimum support, the minimum support is reinterpreted as shown in equation 3.1 to ensure that unreported target behavioral profile are those ones that would never be used even if reported because combining those partial profile to their matches would result in a complete target behavioral profile which does not meet the minimum support.

$$\text{Number of Set Entries} \geq \begin{cases} 0, & \text{for } M < le\ 50\% \\ (MT) - (\frac{1}{2}T), & \text{for } M >\ 50\% \end{cases} \tag{3.1}$$

This equation means that, given a historical period of T time slots and a minimum support M as percentage of the historical period, any complete target profile whose behavioral buffer has at least MT set entries meet the minimum support. Since the historical period is divided into halves, the sensor nodes profiling the target would consider target's behavioral buffer with at least $\frac{1}{2}(MT)$ set entries. But, this condition is not enough because of the implied assumption that the total number of each targets activities is distributed equally over the two half of the historical period. This is clearly not the case since we have already assumed that targets' activities are randomly distributed. A more complete condition is to assume that the other sensor node may have from 0 to $\frac{1}{2}T$ set entries. Recall that there are two sensor nodes monitoring each target, one sensor nodes for each half of the historical period. If the number of set entries is 0, then the complete profile can only meet the minimum support of less

than 50%. Whereas, if the number of entries is $\frac{1}{2}T$, the complete profile will meet at least a minimum support of 50% and meet any minimum support beyond provided the other sensor node has at least $(MT) - (\frac{1}{2}T)$ set entries. Therefore, if sensor nodes report using these two conditions, the sink will be able to construct all the complete target behavioral profiles that meet the minimum support.

Two important remarks must be made here. First, if only one partial profile is reported to the sink, the sink assumes that the complete target profile does not meet the minimum support. However, the sink does not know whether that is true or not because the absence of a partial target profile may be a result of communication problems too. In practical situation, the sink would be sent a notification message about the existence or not of any recorded partial profile. Second, equation 3.1 is transparent to the sink. It is only used at the sensor node level when deciding whether to report recorded targets' activities because sensor nodes have only access to incomplete targets data. In other word, this equation does not change the minimum support level set by the sink. It is only for implementation purpose, although the sink may be aware of it if the user of the network wishes so.

A better alternative to Schedule-Buffer based Data Preparation Mechanism would require the collaboration of sensor nodes grouped in different sponsor sets in order to ensure that both sensor nodes report when necessary. If the sensor node in the first sponsor set reports a target behavioral profile, so should the sensor node in the second sponsor set. This way the two sensor nodes would make sure the complete profile obtained by combining their two partial profiles meet the minimum support before reporting. In the end collaboration would result in even less message transmission.

3.5.3 Fused-Schedule-Buffer based Data Preparation Mechanism

The Fused Schedule-Buffer data preparation mechanism is an improved version of the Schedule-Buffer based data preparation mechanism. The improvement consists in allowing sensor nodes to collaborate with one another during data preparation. The collaboration is of simple form: each sensor node sends recorded partial target behavioral profiles to the one scheduled next and so on until the complete profile is obtained for the desired period of time. Algorithm 3.5.11 gives a formal description for this mechanism.

Algorithm 3.5.11 Fused-Schedule-Buffer based Data Preparation

Upload data parameters;
while Network is Active **do**
  for Each Sensor Node **do**
    Target Monitoring;
    Report Target Activity;
  end for
  for Each received message M **do**
    Extract Activity Set AS_{T_i} ;
    Insert { AS_{T_i}, TD };
  end for
end while

In addition to the same assumptions used to define Schedule-Buffer based Data

Preparation Mechanism, we assume the existence of direct collaboration between sensor nodes in different sponsor sets. This means that not only does each sensor nodes in a sponsor set know which targets whose activities it must monitor, record, and report to the sink but also each sensor nodes knows first hand what other sensor nodes (in a different sponsor sets) cover the same targets. Collaboration ensures that sensor nodes are provided with more information about the targets during the lifetime of the network for better performance. In facts, the sensor nodes progressively learn more about the monitored targets. In this case, sensor nodes learn about targets' past behavior before they begin to profile that same target.

This learning is done as a handoff mechanism at the end of the sensor nodes' scheduled time. When the end of the scheduled time is reached, one of the following two options is selected: The first option is used when the end of the scheduled time does not correspond to the end of the historical period. In such a case, the sensor node sends all of its target buffers' content to the next-sensor node scheduled to take over the monitoring activities of the same targets. However, target buffers' contents may be sent to one or several sensor nodes depending on how many sensor nodes are in a sponsor set. During this process, sensor nodes again use equation 3.1 to find out whether it is appropriate to send. The second option is used when the end of the scheduled time corresponds to the end of the historical period. At this moment the target buffers contain the activity record of all the targets for the complete historic period. Therefore, the scheduled sensor nodes can prepare and send the complete behavioral profiles to the sink. This time, the sensor nodes do not need to use equation 3.1 when reporting because only sensor nodes scheduled as stated in the first option use that equation to determine whether recorded targets' profiles would

meet the minimum support.

The implication for our experiments is that only those sensor nodes from the sponsor set responsible for monitoring the second half of the historical period will report the targets' activities to the sink. This is because we only have two sponsor sets, each one scheduled for half of the historical period. The targets' activities recorded by the first sponsor set, which is responsible for monitoring the first half of historical period, are sent to the sensor nodes in the second sponsor set at the end of the first half of the historical period. Remember that sensor nodes from each sponsor set know in advance the sensor nodes from the other sponsor set which monitor the same set of targets. Then sensor nodes in the second sponsor set combine the partial target behavioral profiles received from the first group of sensor nodes with their recorded ones to obtain the complete target behavioral profiles. At the end of the second half of the historical period, the sensor nodes in the second sponsor set must report the targets' activities detected throughout the entire historical period provided the minimum support is met.

Algorithm 3.5.12 Upload Data Parameters

Sink Node:

Broadcast global schedule;

Broadcast mining parameters $\{ T_{hist}, \lambda, min\_sup \}$

Further improvement to both Schedule-Buffer and Fused-Schedule-Buffer based Data Preparation Mechanisms is possible if the scheduled sensor nodes are allowed to predict whether or not the complete profile will meet the minimum support requirement. In the case the scheduled sensor node predicts that the next-scheduled

Algorithm 3.5.13 Target Monitoring

Sensor Node:;

Upon receiving mining parameters and schedule Time;

while Healthy **do**

if Scheduled **then**

Given Set of Covered target $T = \{T_a,\ T_b,...\ T_i\}$ respectively

Then $S_{next} = \{S_a,\ S_b,...\ S_i\}$ is set of the sensor node next-scheduled to monitor target T;

Observe Covered Targets;

end if

if Not Scheduled **then**

Go into Sleep Mode;

end if

end while

sensor node will meet the minimum support, then it would send its partial profile to the next-scheduled sensor node. Otherwise the sensor node will not send the partial profile of the target, instead it will notify the next-scheduled sensor node to either not bother sending its own partial profile or to start over the partial profiling with new mining parameters.

3.5.4 Generalization to Multiple Sponsor Sets

So far we have assumed the number of sponsor set to be limited to two sponsor sets. However, if any WSN is expected to last several years after the initial deployment and if throughout that lifetime several sensor nodes are continuously replaced by new

Algorithm 3.5.14 Observe Covered Targets

T_{Sche} = Scheduled Time; Time = Current Time;

$T = SetofCoveredTarget$;

if Waking up **then**

Upon Receiving M from previous sensor nodes;

AS($T_i j$ = AS(T_i) extracted from M.

end if

for *Each* $T_i \in T$ **do**

B(T_i) = Behavioral Buffer of Target T_i;

end for

while $CurrentTime \leq Time + T_{Sche}$ **do**

if event is detected **then**

Slot Number = $CurrentTime/\lambda$)

Let T_i be the target where event is detected and B(T_i) is the corresponding buffer;

if $B(T_i)[slotNumber]$ not set **then**

Set $B(T_i)([slotNumber])$;

else

Do nothing;

end if

end if

end while

Report Target Activity;

ones, it is thus normal to make new sponsor sets to include the newly deployed sensor nodes. Hence, the number of sponsor set may be significantly higher than two during the network lifetime. An alternative is to have one sponsor set in which failing sensor

Algorithm 3.5.15 Report Target Activity

Sensor Node:;

Require: End of scheduled time reached

if Current Time = Time + Historical Period **then**

for each target T_i within the sensor's coverage **do**

AS(T_i) = The set of time slots that have set bits in B(T_i) + AS(T_{ij});

if cardinality of AS(T_i)$\geq$ min_sup **then**

M = { Sensor id, T_i, $AS(T_i)$ };

Send M to the Sink;

end if

end for

else

for Each Target T_i within sensor coverage **do**

AS(T_i) = The set of time slots that have set bits in B(T_i) + AS(T_{ij});

M = { Sensor id, T_i, $AS(T_i)$ };

Send M to S_i i.e next-scheduled sensor;

end for

end if

nodes are removed and new ones added. In practical cases, each sensor node would be scheduled for a very small period compared to the historical period needed to adequately profile any target.

Equation 3.1 was not valid for minimum support less than 50% because the scheduled time was 50% of the historical period. In order to be valid for lower minimum support levels, the scheduled time has to be significantly smaller that the historical period. Therefore, for a historical period of T_{Hist} time slots and the sensor node is

scheduled for time T, the equation 3.1 becomes valid for minimum support M greater than $\frac{T}{T_{hist}} * 100\%$ then equation 3.1 becomes

$$\text{Number of Set Entries} \geq \begin{cases} 0, & \text{for } M < le\ \frac{T}{T_{hist}} * 100\% \\ (MT) - (\frac{1}{2}T), & \text{for } M > \frac{T}{T_{hist}} * 100\% \end{cases} \tag{3.2}$$

The smaller the scheduled time is, compared to the historical period, the smaller the minimum support level that can be used with the above equation.

3.6 Proposed Solution Limitations

Before we discuss the implications of our research, it is appropriate to mention its limitations.

The data preparation mechanisms proposed here may not be generalized to all WSN topologies where data mining is to be carried out. Furthermore, we only explore the energy aspect of data preparation and we have assumed this aspect does not affect the results of data mining as such only a framework of Target-based association rules has been proposed thus the actual data mining is not part of this thesis.

Another limitation is the way sponsor set have been used in the experiment. It is assumed that all sensor nodes in the same sponsor set stop working almost at the same time which is unlikely in practical cases because each sensor node perform its duty at different rate (different target activity rate and different number of covered targets). Instead of having a fixed number of sponsor set, it is wiser to have one sponsor set in which sensor nodes are added to replace failing or failed ones. The

proposed mechanisms would then concern only those sensor nodes to be replaced and the new ones.

Despite these limitations, this study will demonstrate the impact of the proposed data preparation mechanisms on the network performance.

3.7 Summary

Target-based Association Rules is a Knowledge Discovery technique designed specifically for POCW. In contrast to Sensor Association Rules, Target-based Association Rules discovers the correlation between the set of targets covered by the network. The activity set of each target is extracted from the sensor nodes covering that target. Three mechanisms for preparing the data needed for generating these rules have been introduced, namely: the All-Nodes, the Schedule-buffer, and the Fused-Schedule-buffer data preparation mechanisms.

Chapter 4

PERFORMANCE EVALUATION

This chapter provides the performance evaluation of the data preparation mechanisms used for mining Target-Based Association rules data miner.

The chapter is organized as follows. Section 4.1 introduces the testing methodologies used to evaluate the performance of the proposed solutions. Section 4.2 presents a performance evaluation of the proposed solutions. Section 4.3 summarizes the chapter.

4.1 Solution Testing

4.1.1 Introduction

The solution implementation has been divided into three parts for comparison purpose. The first part is about an existing benchmark that we refer to as *All-Node direct reporting*. The second part is termed *Schedule-Buffer direct reporting*, and the third part which is a modification of the Scheduled-Buffer direct reporting is named *Fused-*

Schedule-Buffer direct reporting. In the All-Node direct reporting, the sensor nodes in a sponsor sets monitor the activities of all covered targets during their scheduled time. Each sensor node report the recorded targets behavioral data following the All-Node data preparation mechanism. In the Schedule-Buffer direct reporting, sensor nodes report the recorded targets behavioral data following the Schedule-Buffer data preparation mechanism. Finally, in the Fused Schedule-Buffer direct reporting, sensor nodes report the recorded targets behavioral data following the Fused-Schedule-Buffer data preparation mechanism.

Direct reporting means that a sensor node is able to send its messages to the sink and no further processing is done while the message is on the way to the sink. We will use wireless connection for all communications between sensor nodes and the sink. For this matter we selected the first order radio model presented in [5]. simulation parameters used in our simulations are summarized in table 4.1 and are used to computed consumed energy during data transmission and handling. A more efficient and practical transmission scheme would involve a multi-hops scheme but this alternative was not considered so far for our analysis.

4.2 Performance Evaluation

To prove that our solution solved the problem defined in the research problem statement section, we now present the performance evaluation of the proposed data preparation mechanisms: this section presents a comparison of the performance of the POCW during the data preparation process using All-nodes, Schedule-buffer and Fused-Schedule-buffer based data preparation mechanisms.

Recall that the main goal of the proposed algorithms is to improve energy efficiency of the KDD process in the context of WSN. We know communication consumes most energy and we set out to define mechanisms that limit these costs as much as possible. Of the entire KDD process, one particular phase drain most energy at the sensor level. this is the transmission of data prior to data processing (data mining) at the end of data preparation in the pre-processing phase. We simulated three different mechanisms as described in the previous section and we now present their performance.

We have selected simulation parameters according the intended application of POCW which require deployment over large area hence the presence of a large number of targets and sensor nodes, because the sensing range may vary from a very short distance to long distance we have also varied the sensing range of the used sensor nodes. The sampling rate (slot size) was selected to be a minutes for tractability and the energy parameters were from commonly used components of sensor nodes.

The metrics used to evaluate the performance of the network are: the number of messages needed to report the activity sets to the Sink, and the average energy consumptions per sensor node. We considered two main sources for energy consumption: transmission energy and energy cost to maintain storage buffer at the sensor node level (i.e., the energy consumption for read, write and erase operation) as listed in Table 4.1. The metrics were measured from the simulation of the three data preparation mechanisms using the simulation parameters listed in table 4.1. We assume that all the sensor nodes have the same sensing range R and transmission range T, where $T \geq R$ and R is greater or equal to the average distance between two consecutive targets to ensure overlap in sensor covered targets. Each node uses its full sensing

Table 4.1: Simulation parameters used in the performance evaluation of the proposed three data preparation mechanisms. The energy parameters were selected from components commonly used to build sensor nodes

Parameter	Value
Grid Area	500×500
Number of Sensor	500, 1000, 5000
Number of Targets	64, 128, 320
Sensing Range	15, 63, 150 m
Minimum Support	10%, 50%, 90%
Historical Period	5 Days
Slot Size	60 sec
Message Size ??	480 kbits
Read, Write, Erase from flash ??	0.017μJ/Byte
Transmission, Receive Energy ??	50 nJ/bit
Transmitter's Amplifier ??	100 pJ/bit/m^2

coverage for target monitoring, but its effective sensing range is limited to a fraction of its maximum transmission range. For target activity prediction, we assume that target observations would obey a Poisson distribution [24]. However, it is possible to modify these parameters depending on applications and sensor specifications respectively. Three different minimum support, 0%, 50%, and 90% minimum support, were used to transmit the sensed data to the sink. each sensor was then expected to report

only if detected activity sets meet the minimum support criteria.

Targets are assumed to be located at fixed points in a square grid of 500×500 meters. 64, 128, and 320 targets have been used and 500, 1000, 10000 sensor nodes deployed randomly. To evaluate the energy consumption of these three mechanisms, we have used the first order radio model introduced in [5]. According to that model, equation 4.1 estimates the energy consumption for sending a k bit message across distance d meters. E_{elec} is the energy consumption used to run the transmitter and the receiver which is estimated to be 50 nJ/bit **??**. E_{amp} = 100 pJ/bit/m^2 is the energy consumption for the transmitter's amplifier **??**. In our experiments, we have assumed that k = 480 bits and d was varied from 15m, 30m and 63m. Equation 4.2 shows the energy consumption of running the receiver circuit. In addition to the transmission and receiving costs, there is an extra cost of running the storage devices used with Schedule-buffer and Fused-Schedule-buffer mechanisms. We have assumed the use of a Toshiba 16MB NAND flash memory that costs 0.017μJ to read, write, and erase a byte of data [43, 44].

$$E_{TX}(k, d) = E_{elec} * k + \varepsilon_{amp} * k * d^2 \tag{4.1}$$

$$E_{RX}(k) = E_{elec} * k \tag{4.2}$$

The performance evaluation was divided into three experiments, One for each data preparation mechanism, essentially each experiment achieves the same objectives but using different data preparation mechanisms and simulation parameters.

The first experiment simulated the All-Nodes data preparation mechanism. The second experiment simulated the Schedule-Buffer data preparation mechanism. Fi-

nally, the third experiment simulated the Fused-Schedule-Buffer data preparation mechanism. The network sensor nodes are divided into two sponsor sets. Each sponsor set is then scheduled for only half of the total historical period. The first sponsor set group is scheduled for the first half of the historical period, and the second sponsor set is scheduled for the second half.

In the All-Nodes data preparation mechanism simulation. All sensor nodes of the scheduled sponsor set monitor all targets within their coverage. Once an activity is detected at one of the targets within coverage, the responsible sensor node immediately reports that activities to the sink. The overall number of messages sent to the sink node were computed on daily basis while the average energy consumption per node were computed at the end of the historical period. The second and the third experiment were different from the first in the operating mode. The difference is that sensor nodes used in the second and the third simulation are assumed to have storage buffers. The buffers allows the sensor node record target activities and to perform data preparation on these data. Sensor nodes only monitor a specific set of targets (the selection of which is done by the sink) and data preparation and transmission to the sink are performed at the end of the scheduled time. The third simulation differs from the second in the way sensor nodes report detected activities to the sink. In the second experiment, all sensor nodes report the recorded target activities to the sink while in the third experiment only those in the second sponsor set report to the sink.

The target activities detected by sensor nodes of the first sponsor set are sent to the sensor nodes of the second sponsor set at the end of the first half of the historical period. Remember that, sensor nodes from each sponsor set know in advances the sensor nodes from the other sponsor set which monitor the same set of targets. The

sensor nodes in the second sponsor set combine the activity records received from the first group of sensor nodes together with their own to get the full historical period. Then at the end of the second half of the historical period, sensor nodes of the second sponsor set report the target activities detected throughout the entire historical period provided the minimum support is met. This mechanism is expected to further improve the network performance. The proof comes from the Cauchy-Schwartz inequality in linear algebra which can be interpreted in this case as two forms of learning. In this case sensor nodes learn about the past target behavior before they begin to profile that same target too. This learning is done as a handoff mechanism at the end of the scheduled time.

The simulation program was done using *Matlab 2007* because of familiarity with this software package. A description of the WSN to be used with all the required parameters and tasks that must be performed on the WSN are specified in a script file named *Main Program*. Structure array class built in Matlab was used to build the WSN network structure which is composed of 3 layers. This structure allows easy manipulation of WSN parameters. The first layer is the WSN itself and includes the sink node, the second layer is the collection of sensor networks nodes and the third layer is the collection of the targets nodes.

We created a *Main Program* that consists of several other small Matlab functions used to create the wireless sensor network first, then perform all the tasks by calling different functions. Within *Main Program*, three identical WSN are created with the same parameters (same number of sensor nodes and targets, same level of energy reserve, all nodes and targets have the same spatial positioning in the network, sensor nodes with same names in each of the networks covers the same set of targets, the

only difference is the length of their scheduled activity period which consequently determine the extent of their historical record).

The main program builds a WSN by specifying all required parameters. The following parameters were given to implemented WSN: the name of the WSN network, the width and length of the WSN coverage area (Field Size), the number of sensor nodes, number of targets nodes, the number of sink nodes, the node structure, the sink structure and the target structure, and sensor nodes maximum transmission range. The effective transmission range is a fraction of the maximum transmission range of the sensor node.

The sensor nodes that constitute the WSN are stored as a structure array of sensor nodes. It is assumed that all sensor nodes have identical physical characteristics i.e they have the same transmission range, battery capacity, antennas types, and modulation/coding schemes. The main program also specifies the following fields for each node: *power Level, position, sensor Name, messages, covered Targets, message Count, buffer.* The *power Level* field reflects current battery reserve as percentage of full charge. The *position* field contains the coordinates of the node in the coverage area. The *sensor Name* field is the identifier of a node. The *messages* field contains received external messages. The *covered Targets* is the set of targets within sensing range. The *message Count* tracks the number of each covered target activities. Lastly, *buffer* contains covered target activity sets.

The target nodes that constitute the WSN are also stored as a structure array of

target nodes. No assumption about the target nodes physical parameters is made but for simulation purpose each target is given a set of parameters that it must broadcast to all of its covering sensor during the detection process. Therefore, the main program also specifies the following fields for each target node: *power Level, position, covering Sensor, target Name, events Record, number Of Events, message Count, event Rate.* The *event record* is the event occurrence time in chronological order for the historical period. *number Of Events* is the number of events that occurred during the historical period. *message Count* is the number of message that are sent out to each covering sensor during the historical period.

The above parameters are passed on to *sensor Network Set Up* function which return a WSN that meets these specifications. This function initialize the WSN structures by deploying in the targets nodes, then the sensor nodes, and finally the sink node.

The next step is then target event detection and reporting processes. To ensure that we have the same target activity sets in all the three experiments. The target activities record are created separately then sensor nodes in the each of the WSN networks are triggered to detect their covered target activity sets. The detection process and the recording process is different for all three networks. The difference results from the historical period used by each sensor nodes in each networks.

Each one of the targets is assigned a random number of events, and these events are generated with a pseudo random numbers generator, then the event rate fitting

this set of data is computed. Once all of the targets are assigned their activity record, the covering sensor node can detect event generated by the targets in their coverage area and store this event record in the corresponding target behavioural buffer.

The process of detecting targets activities is done in two steps. The first step is event triggering and the second one is event detection and recording.

During event triggering each target send a message to all of its covering sensors. The sent message contains the event occurrence time and the target name. To avoid mistakes or have differences in recorded target activities from WSN to WSN, event triggering was done from a single set of target event record as follow. In the case of

All-Node direct reporting: sensor nodes are sent copies of the target activity record (event are ordered from earliest to latest) covering the entire historical period.

Sponsor-Set direct reporting: sensor nodes are sent only partial copies of the target activity records. Events that happen prior to a time T (T being half of the entire historical period) are sent to sensor nodes in the first sponsor set while the event happening post time T are sent to the second sponsor set nodes.

Modified Sponsor-Set direct reporting: is similar to Sponsor-Set direct reporting; however, the second sponsor set nodes are sent the complete target event record covering the entire historical period.

The second step of the process of detecting targets activities is the event detection and recording. sensor nodes detect the event that just occurred by performing two operations. The first operation which is the *event detection* is the target message reception and the second operation which is the *event recording* is a save operation to record the event in the corresponding target activity set. In reality, every time a sensor detect an event during any given time slot, it would have to check whether any event has been already detected from that target previously but within the same time slot. This is equivalent to a memory read operation where the sensor nodes check the buffer entry for the current time slot to see if it is set. This is so that the sensor does not perform unnecessary save operation if the buffer entry corresponding to the time slot is already set. However, this is not what is done in the simulation, instead of read operation a save operation is performed each time an event is detected. This is equivalent to overwriting the current buffer slot because the save operation cost the same as the read operation in terms of energy. To end event detection, the sensor energy reserve level is adjusted to account for the performed operations.

As mentioned before, event are directly reported to the sink without any routing. However, there are differences in event reporting as a result of data preparation mechanism used. Once the events are generated and detected the the nodes must report the detect activities to the sink according to the predetermined support level. this assign a name, number of sensor and number of targets.

The report message contains the reporting sensor name, the reported target's name, and the target behavioural buffer content. Also every time a sensor node send a message it performs two actions: *one read* operation to read from the buffer, and

one transmission operation to send the message. These two operations have different costs in term of energy. The sensor energy level are then adjusted accordingly until all messages have been sent.

We have compared the total number of transmitted message and the average energy consumption per node for POCW of 1000 sensor nodes for a 5 days historical period.

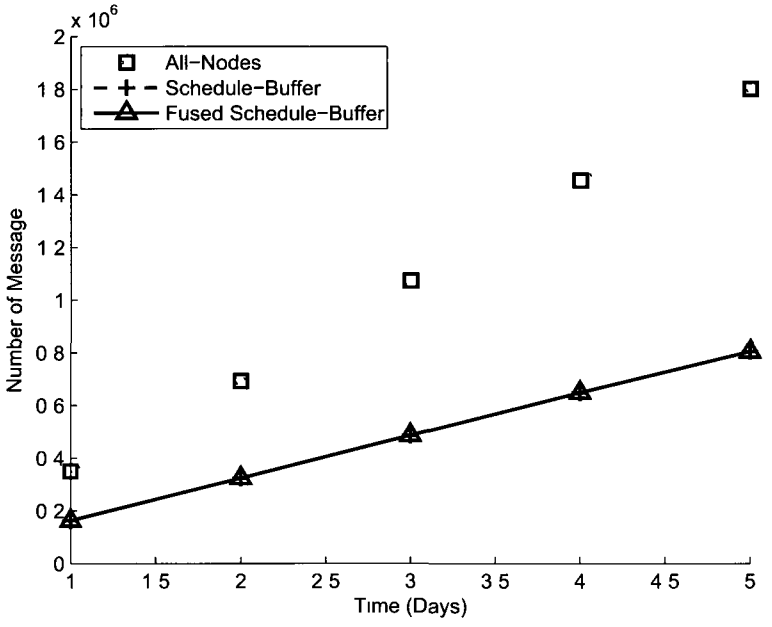

Figure 4.1: Total Number of Messages at 0% Minimum Support.

Figures 4.1, 4.2 and 4.3 show the total number of messages needed to report the activity sets to the Sink at minimum supports values of 0%, 50% and 90% respectively while figures 4.4, 4.5 and 4.6 show the energy consumption per node for the same minimum support levels.

We can see, from these figures, that the total number of transmitted messages for

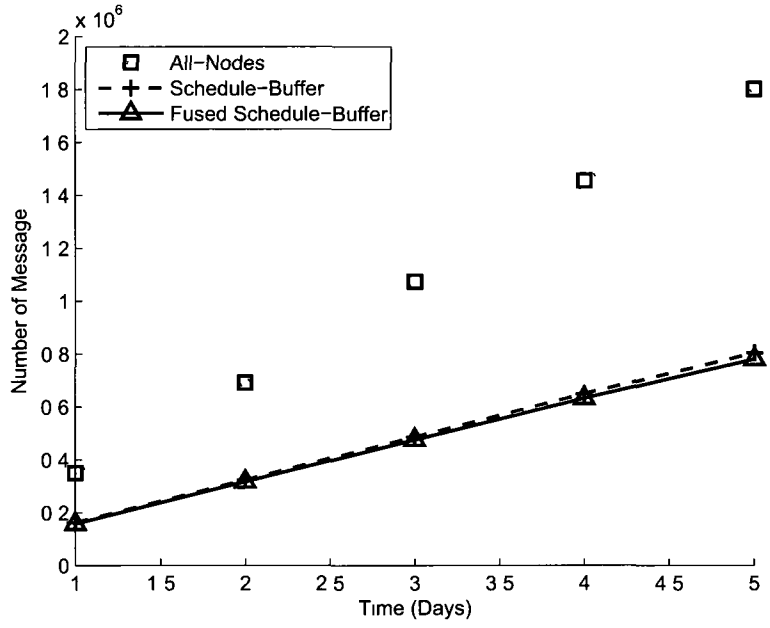

Figure 4.2: Total Number of Messages at 50% Minimum Support.

Fused-Schedule-buffer mechanism is less than the total number of transmitted messages using Schedule-buffer mechanism and significantly less than number of messages transmitted when using the All-Node mechanism. If we compare the performance of sensor nodes that use the Schedule-buffer and the Fused-Schedule-buffer mechanisms to the ones using the All-Nodes mechanism, we can see that those sensor nodes using the Schedule-buffer mechanism to report target activities, require only 43-47% of the messages used by the All-Nodes mechanism. The same node using the Fused-Schedule-buffer mechanism achieves an even more significant reduction since it only requires between 35-45% of the total number of messages transmitted if it were to use the All-Node mechanism. This reduction on the number of transmitted messages affects the energy consumption levels directly. The energy consumption per node for Fused-Schedule-buffer mechanism is less than the average energy consumption

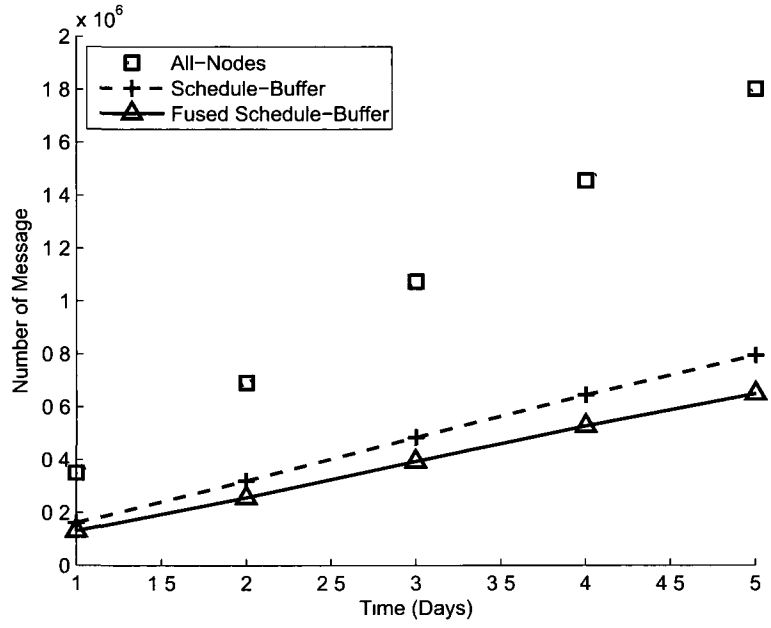

Figure 4.3: Total Number of Messages at 90% Minimum Support.

using Schedule-buffer mechanism and significantly less than energy used in All-Node mechanism. In Fused-Schedule-buffer mechanism, each node that reports activities to sink requires between 42-50% of the energy exhausted by a sensor node using the Schedule-buffer mechanism, and between 30-32% of the energy exhausted by a node in the All-Node mechanism. Note that the Average energy consumption at high minimum support may be very small in the early days if it is mainly spent to maintain the targets' buffers since the activity sets reported to the sink are not that frequent because the minimum support is not met. The reduction in the number of messages is explained by using the global schedule and the buffer mechanisms used by Schedule-buffer scheme. The Schedule-buffer mechanism also outperform the All-Nodes mechanism in terms of energy consumption reduction achieved during data preparation in this simulation the Schedule-Buffer mechanism uses 70-75% of the

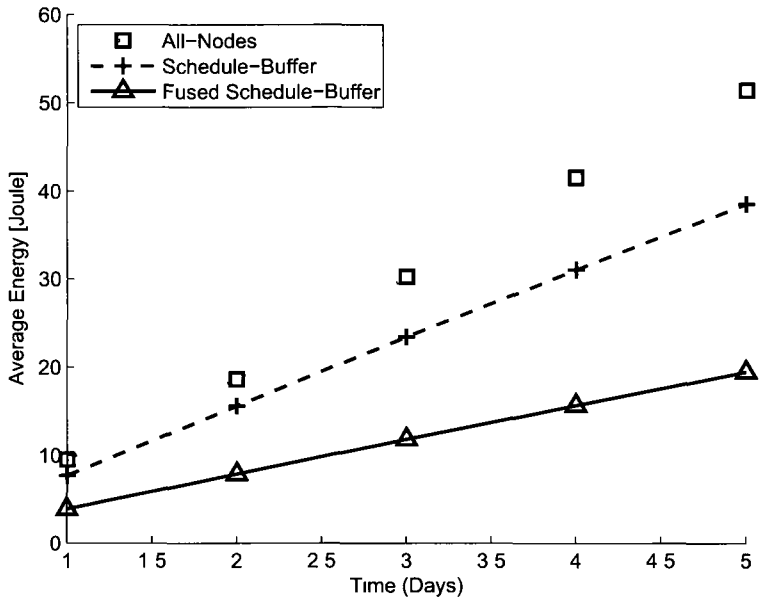

Figure 4.4: Average Energy Consumption at 0% Minimum Support.

energy used by the All-Node mechanism.

Why is the number of transmitted message for the All-Nodes data preparation mechanism higher than the other two mechanisms? The reason is that, when sensor nodes use the All-Nodes mechanism to report, each sensor node transmit a message to the sink every time an new target activity is detected. However, when the same sensor nodes use the Schedule-buffer or Fused-Schedule-buffer mechanisms to report target activities minimize redundancy for two reasons. First, each target is monitored by a single sensor node at any time. Second, the scheduled time is divided into time slots, and all activities observed within the same time slot are recorded as one in the target buffer.

Why do the Schedule-buffer mechanism and the Fused-Schedule-buffer mechanism have the same number of sent message at 0% support value? At 0% minimum sup-

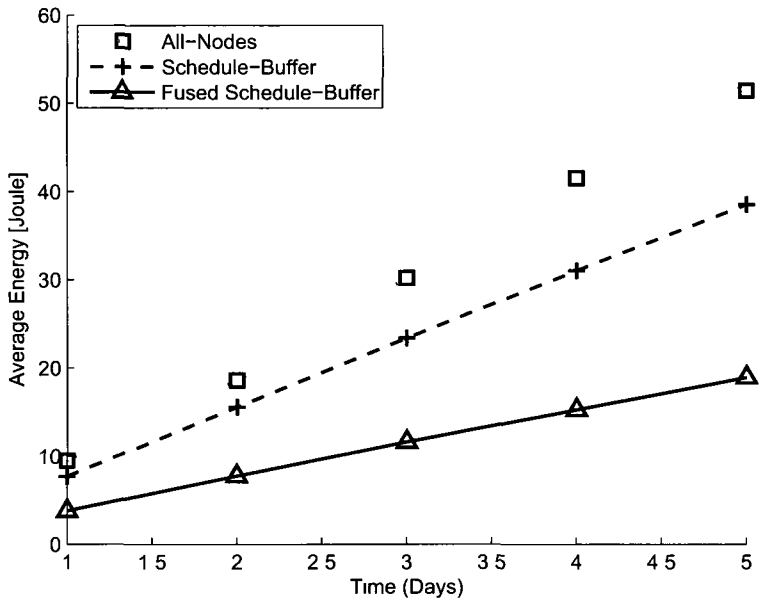

Figure 4.5: Average Energy Consumption at 50% Minimum Support.

port, the Schedule-buffer and Fused-Schedule-buffer always have the same number of transmitted messages because every buffer content is reported. At higher support level, the number of transmitted messages differs because there are different level of optimization when sensor nodes use the Schedule-buffer and the Fused-Schedule-buffer mechanism. In the case of Schedule-buffer mechanism, sensor nodes prepare target behavioral profile from a partial target activity data and want to minimize the worst case scenario in absence of sensor nodes collaboration (i.e having an unreported target profile when it should) hence, more data are transmitted than necessary. In the case of Fused-Schedule-buffer mechanism, sensor nodes can optimize the data preparation process because the complete target activity record is available as a result of sensor nodes collaboration.

These results are more interesting if we compare the performance of each mech-

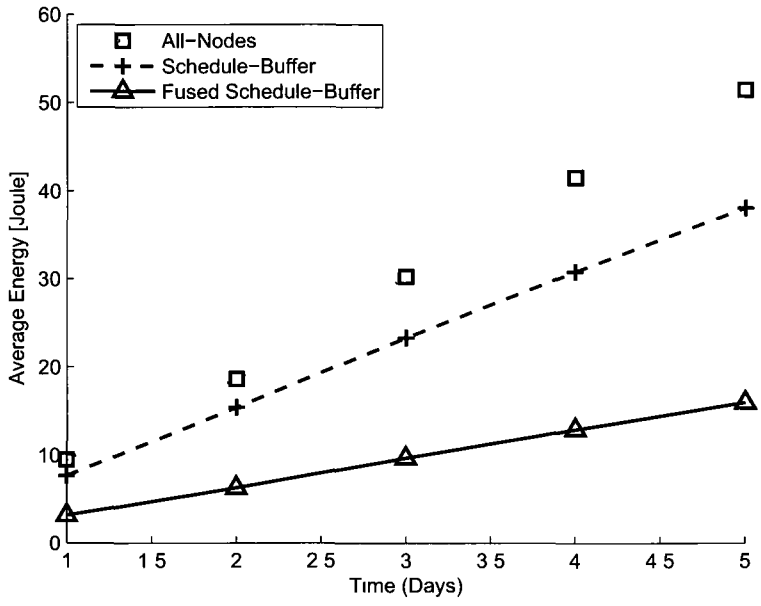

Figure 4.6: Average Energy Consumption at 90% Minimum Support.

anism as a function of the minimum support level, a visual comparison is shown in figure 4.7, 4.9, 4.8, 4.10. The performance of the Fused Schedule-buffer data preparation mechanism improves with increasing minimum support. Notice that it is possible in the case of Schedule-Buffer based mechanism to have the same number of transmitted messages for all minimum support. This is the case when some target that are of interest are very active during some time interval. This is misleading in the sense that the sensor nodes have a partial view of the target profile and may perceive an otherwise inactive target, if viewed over the entire historical period, to be very active. This occurrence is unlikely in the Fused-Schedule buffer because all the data is combined at the sensor node level and not two targets will likely have the same level of activities over the entire historical period. Similar results were also observed independently of the sensing range used.

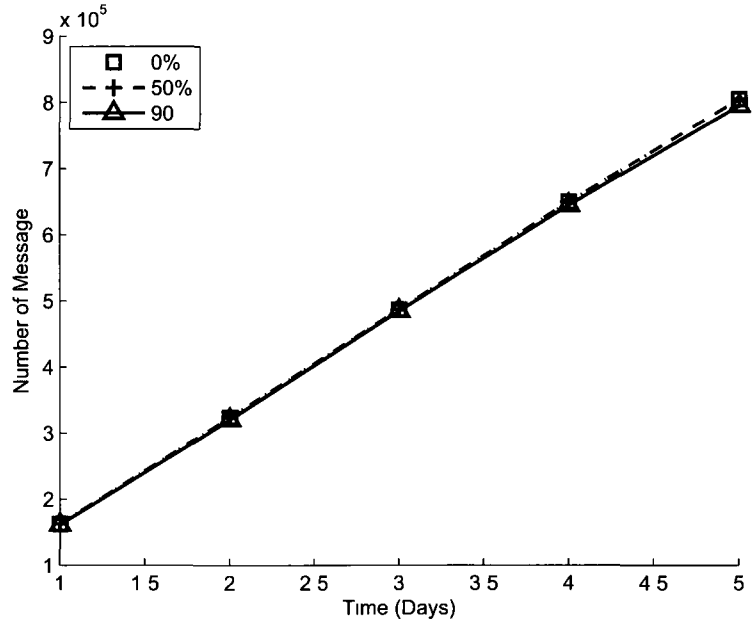

Figure 4.7: Comparing the Total Number of Transmitted Messages when Schedule-Buffer based data preparation mechanism is used with different minimum support levels.

The Schedule-Buffer data preparation mechanism yields the same number of transmitted messages and average consumed energy for minimum support levels between 0% and 50%. This happens because the two sponsor sets are exactly scheduled for half of the historical period. Figure 4.7 (figure 4.9) suggests that the number of transmitted messages (average consumed energy) is the same for all minimum support levels. This is misleading because a zoom of figure (see 4.11 and 4.12) on the third day shows a slight improvement in the number of transmitted messages at 90%. In fact, any minimum support level greater than 50% generates better result although the difference is minor because all of the monitored targets are very active. The reason is that almost all targets are very active. Remember, that the sensor nodes using this mechanism try to minimize the worst case scenario by sending more

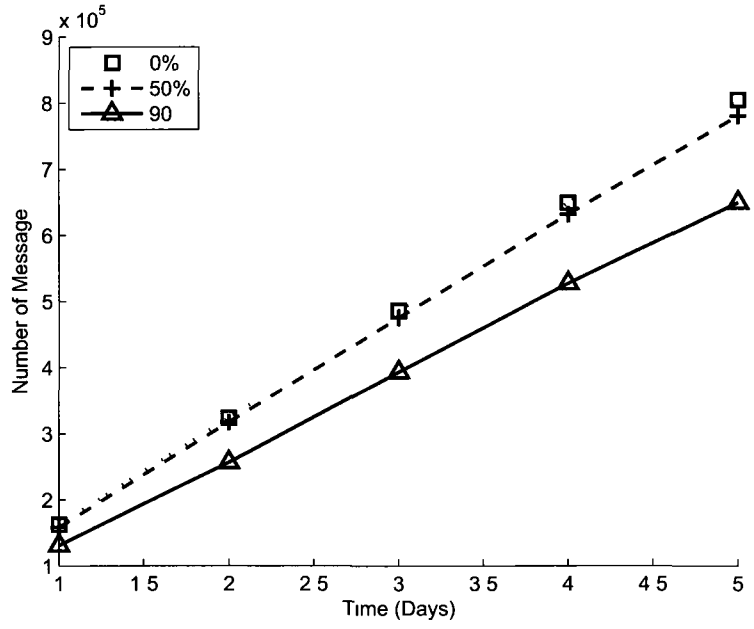

Figure 4.8: Comparing the Total Number of Transmitted Messages when Fused-Schedule-Buffer based data preparation mechanism is used with different minimum support levels.

data than necessary. Hence, they may look the same on the figure but are actually slightly different.

Increasing the number of target nodes in the covered field from 64 to 128 then to 320 did not affect the performance of the proposed data preparation mechanisms. Similarly, there is no noticeable effect when the number of sensor nodes in the network is increased from 1000 to 2000 then 5000 sensor nodes again for a 5 days historical period. However, changing the sensing range did affect the performance of the proposed data preparation mechanisms in an interesting way. The performance was greatly improved and if the sensing range was doubled then the number of message transmitted and the energy cost was halved i.e if the sensing range is multiplied by a factor then the number of messages and the energy cost associated to it is divided by almost the

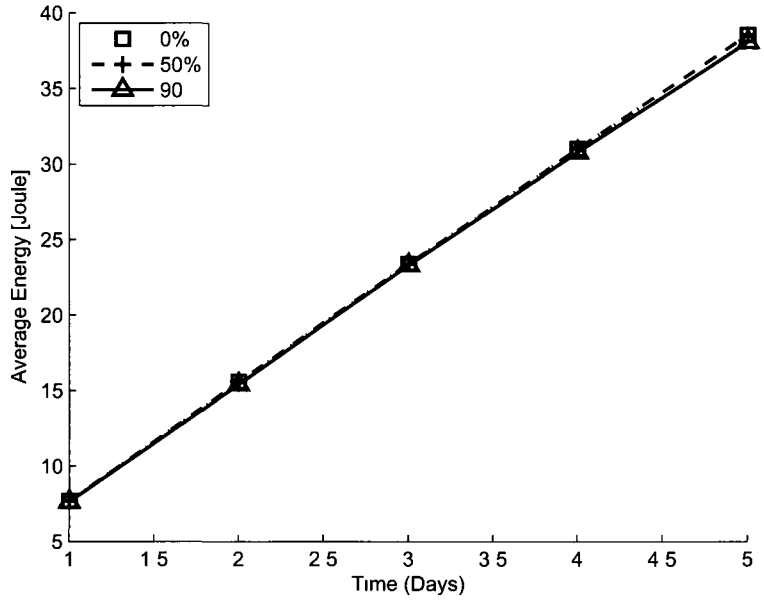

Figure 4.9: Comparing the Average Energy Consumption of Schedule-Buffer based data preparation mechanism at different minimum support levels.

same factor. The difference between the performance of the All-Node mechanism, Schedule-Buffer mechanism and Fused-Schedule-Buffer mechanism respectively, for support values of 0%, 50%, and 90% was still significant but overall the performance is greatly affected by the sensing range of individual sensor nodes. For example, when the sensing range is multiplied by ten, results show that Fused-Schedule-Buffer mechanism still requires fewer messages to extract data than Schedule-Buffer mechanism scheme and significantly less than All-Node mechanism. Schedule-Buffer mechanism requires at least 4% to 5% of the number of messages needed for preparing the behavioral data needed for All-Node mechanism while Fused-Schedule-Buffer mechanism requires only 2% to 3%. The same trend is observed for average energy consumption. Schedule-Buffer mechanism requires at least 9% to 10% of the average energy

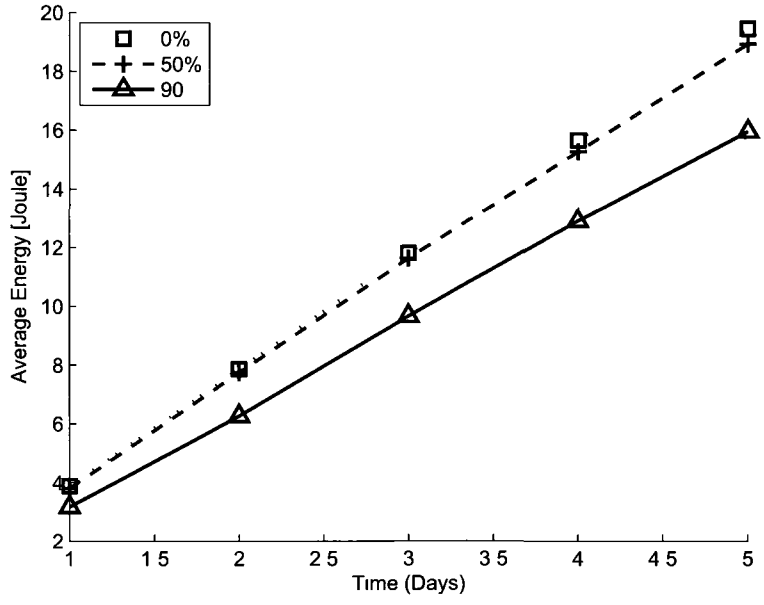

Figure 4.10: Comparing the Average Energy Consumption of Fused-Schedule-Buffer based data preparation mechanism at different minimum support levels.

needed for preparing the behavioral data needed for All-Node mechanism while Fused-Schedule-Buffer mechanism requires only 3% to 6%.

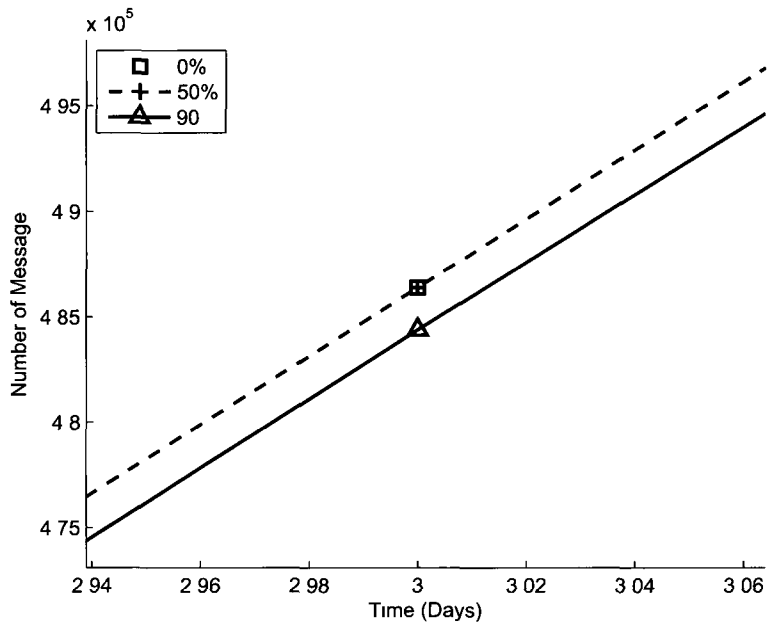

Figure 4.11: Zoom of figure 4.7 on the third day comparing the Total Number of Transmitted Messages when Schedule-Buffer based data preparation mechanism is used with different minimum support levels.

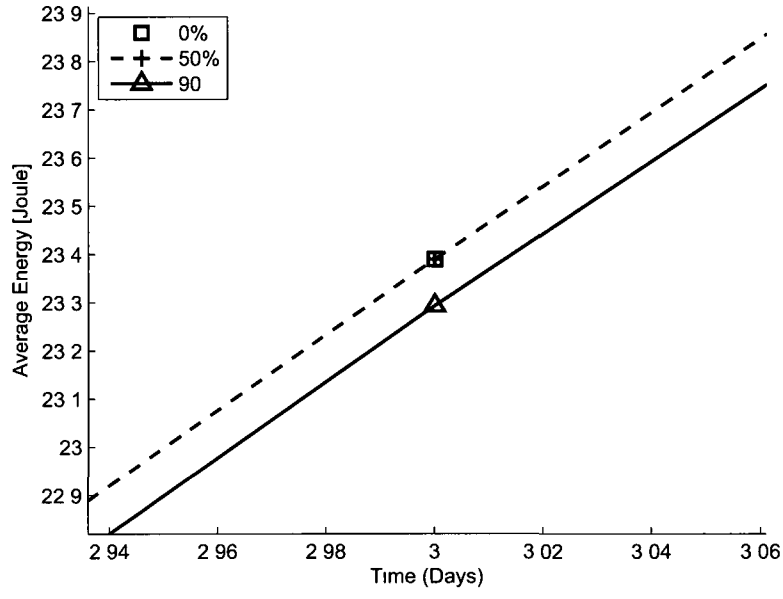

Figure 4.12: Zoom of figure 4.9 on the third day comparing the Average Energy Consumption of Schedule-Buffer based data preparation mechanism at different minimum support levels.

4.3 General Summary

Target-based Association Rules is the Knowledge Discovery technique designed specifically for POCW. In contrast to Sensor Association Rules, Target-based Rules discovers the correlation between the set of targets covered by the network. The activity set of each target is extracted from the sensor nodes covering it. Three mechanisms for preparing the target database needed for generating the Target-based Rules have been introduced namely: the All-Nodes, the Schedule-buffer, and the Fused-Schedule-buffer data preparation mechanisms. Several experiments have been conducted to evaluate these data preparation mechanisms. Results indicate clearly that Schedule-buffer and Fused-Schedule-buffer data preparation mechanism outperforms the All-nodes data preparation mechanism in term of the total number of messages and average energy consumption. Our results show that the proposed approach will not only reduce the energy consumption level in a network but also the energy could allow a better management of the WSN by modifying the operating parameters to deal with the dynamic nature of the network. The proposed solution to find interesting patterns among targets could also be used dynamically to change the sponsor set grouping and further adapt the network resource to current situations.

Chapter 5

THESIS CONCLUSION

Thesis-Conclusion)

At the beginning of this work, we have put forth a number of questions. We feel we are now in a position to answer these questions, and perhaps add a few additional comments to what has already been said.

The main result of this thesis is to have shown that it is indeed possible to have low energy consumption data preparation mechanisms in the context of WSN. That objective has been attained with the proposed data preparation mechanisms achieving between 50% to 70% reduction in message transmission, and between 25% to 80% reduction in energy cost associated with data preparation compared to traditional data preparation methods.

Interesting results were also obtained from a number of different experiments, that we shall discuss here. In one of them, we scaled the WSN with sensor nodes and the simulation results showed no effect on the performance of WSN in terms of the number of transmitted messages and the average energy consumption. The same is

true when the number of targets in the field is increased.

In another experiment, the sensing range was increased to twice the initial value then to five times the initial value. The results show that increasing the sensing range reduces the number of transmitted messages i.e. the energy consumption in the networks. This may be due to the fact that messages exchanged among the sensor nodes are significantly reduced when a few sensor nodes can monitor all the targets effectively. This was rather surprising and further investigations may be needed to interpret this observation fully.

We have shown that Target-based Association Rules is a Knowledge Discovery technique, designed specifically for Point-Of-Coverage wireless sensor networks (POCW). In contrast to Sensor Association Rules, Target-based association rules discover the correlation between the set of targets covered by the network. The activity set of each target is extracted from the sensor nodes covering that target. Three mechanisms for preparing the target database needed for generating the Target-based association rules have been introduced. Simulation experiments have been conducted to evaluate these data preparation mechanisms, and their results indicate clearly that Schedule-buffer and Fused-Schedule-buffer data preparation mechanisms outperform All-nodes data preparation mechanism in term of the total number of messages and the average energy consumption.

We now arrive at the final and most important question, do Target-based association rules for POCW and their associated data preparation mechanisms namely: the All-Nodes, the Schedule-buffer and the Fused-Schedule-buffer data preparation mechanisms serve any purpose?

Our claim is not that our algorithms are always the method of choice, but rather

that under some conditions these algorithms will be the most efficient to use. In fact, we expect wireless sensor networks to adopt similar algorithms in the future. Our results clearly show that the proposed data preparation approaches will not only reduce the energy consumption level in a WSN network but also could allow a better management of the WSN by allowing on the fly changes of the operating parameters in order to deal with the dynamic nature of the WSN networks.

In any case, we feel that one of the most important result of this study is with respect to other type of data preparation mechanisms. The reason is that clearly the result of this study shows a net reduction in energy consumption even though the Target-based association rules may not be applicable to other applications of WSN, we feel that the Schedule-Buffer and Fused-Schedule-Buffer data preparation mechanisms can be adapted to any centrally managed WSN.

These association rules could be useful to track military targets or environmental hazards to name few applications. Applications that have become very important in the last few years and that will increasingly rely on innovative data mining techniques.

The objective of this research was to define new behavioral patterns for Point-Of-Coverage wireless sensor networks and their associated data preparation mechanisms for mining those patterns. That objective has been attained with the presentation of Target-Based Association Rules and their associated data preparation mechanisms, and at least a 50% reduction in the number of transmitted messages and at least a 25% reduction in energy cost, compared to traditional data preparation method, has been achieved in simulations. We, finally, conclude by noting that this study has raised a lot of questions which we intend to pursue further in the future. The most interesting ones are those of how to use these association rules for better WSN

management, and to improve QoS and WSN lifetime.

Bibliography

[1] I. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "A survey on sensor networks," *IEEE Communications Magazine*, vol. 40, no. 8, pp. 102–114, 2002.

[2] F. Zhao and L. J. Guibas, *Wireless Sensor Networks: An Information Processing Approach.* Morgan Kaufmann publisher, 2002.

[3] A. Boukerche and S. Samarah, "A novel algorithm for mining association rules in wireless ad hoc sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 19, no. 7, pp. 865–877, 2008.

[4] K. Akkaya and M. Younis, "A survey on routing protocols for wireless sensor networks," *Ad Hoc Networks*, vol. 3, no. 3, pp. 325–349, 2005.

[5] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient communication protocol for wireless microsensor networks," in *Proceedings of the 33rd Hawaii international Conference on System Sciences*, (Maui, Hawaii, USA), pp. 8020–8030, January 2000.

[6] C. Intanagonwiwat, R. Govindan, D. Estrin, J. Heidemann, and F. Silva, "Directed diffusion for wireless sensor networking," *IEEE/ACM Transactions on Networking*, vol. 11, pp. 2–16, Feb 2003.

[7] S. Madden, M. J. Franklin, J. M. Hellerstein, and W. Hong, "Tag: a tiny aggregation service for ad-hoc sensor networks," *ACM SIGOPS Oper. Syst. Rev.*, vol. 36, no. SI, pp. 131–146, 2002.

[8] S. Samarah, A. Boukerche, and R. Yonglin, "Coverage-based sensor association rules for wireless vehicular ad hoc and sensor networks," in *IEEE Global Telecommunications Conference 2008*, pp. 1–5, Dec 2008.

[9] S. Samarah and A. Boukerche, "Chronological tree-a compressed structure for mining behavioral patterns in wireless sensor networks," *Journal of Interconnected Networks*, Fall 2008.

[10] B. Rajagopalan and M. W. Isken, "Exploiting data preparation to enhance mining and knowledge discovery," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 31, pp. 460–467, Nov 2001.

[11] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining.* Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[12] P. Cabena, P. Hadjnian, R. Stadler, J. Verhees, and A. Zanasi, *Discovering data mining: from concept to implementation (IBM Books).* Upper Saddle River, N.J., USA: Prentice Hall, 1998.

[13] D. Pyle, *Data Preparation for Data Mining.* San Mateo, CA, USA: Morgan Kaufmann, 1999.

[14] A. Boukerche and S. Samarah, "An efficient data extraction mechanism for mining association rules from wireless sensor networks," in *IEEE International Conference on Communications, ICC'07*, pp. 3936–3941, June 2007.

[15] A. Boukerche and S. Samarah, "A new in-network data reduction mechanism to gather data for mining wireless sensor networks," in *Proceedings of the 10th ACM Symposium on Modeling, analysis, and simulation of wireless and mobile systems, MSWiM '07*, pp. 70–77, 2007.

[16] A. Boukerche and S. Samarah, "A performance evaluation of distributed framework for mining wireless sensor networks," in *Proceedings of the 40th Annual Simulation Symposium, ANSS '07*, pp. 239–246, March 2007.

[17] V. S. Tseng and K. W. Lin, "Mining temporal moving patterns in object tracking sensor networks," in *Proceedings of the International Workshop on Ubiquitous Data Management, UDM'05*, pp. 105–112, April 2005.

[18] S. K. Chongl, S. Krishnaswamyl, S. W. Lokez, and M. M. Gaben, "Using association rules for energy conservation in wireless sensor networks," in *Proceedings of the 2008 ACM symposium on Applied computing, SAC '08*, pp. 971–975, March 2008.

[19] S. Ratnasamy, B. Karp, S. Shenker, D. Estrin, R. Govindan, L. Yin, and F. Yu, "Data-centric storage in sensornets with ght, a geographic hash table," *Mob. Netw. Appl.*, vol. 8, no. 4, pp. 427–442, 2003.

[20] S. Tilak, N. Abu-Ghazaleh, and W. Heinzelman, "A taxonomy of wireless micro-sensor network models," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 6, no. 2, pp. 28–36, 2002.

[21] T. S. Rappaport, *Wireless Communications Principle and Practice, 2nd ed.* Upper Saddle River, N.J., United States of America: Prentice Hall, 2002.

[22] C. Song and M. Guizani, "Energy map: Mining wireless sensor network data," in *IEEE International Conference on Communications (ICC) 2006*, vol. 8, (Istanbul, Turkey), pp. 3525–3529, 2006.

[23] M. Cardei and D. Du, "Improving wireless sensor network lifetime through power aware organization," *Wireless Networks*, vol. 11, no. 3, pp. 333–340, 2005.

[24] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering, Third Edition.* Upper Saddle River, NJ, USA: Pearson/Prentice Hall, 2008.

[25] M. J. Sailor and J. R. Link, "Smart dust: nanostructured devices in a grain of sand," *Chemical Communications*, vol. 11, pp. 1375–1383, 2005.

[26] P. Lyman and H. R. Varian, "How much information." http://www.sims.berkeley.edu/how-much-info-2003, May 2003.

[27] U. M. Fayyad, "Data mining and knowledge discovery in databases: implications for scientific databases," in *Proceedings of the Ninth International Conference on Scientific and Statistical Database Management 1997*, (Olympia, Washington, USA), pp. 2–11, Aug 1997.

[28] W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus, "Knowledge discovery in databases: An overview," *AI Magazine*, vol. 13, no. 3, pp. 57–70, 1992.

[29] F. Mhamdi and M. Elloumi, "A new survey on knowledge discovery and data mining," in *Second International Conference on Research Challenges in Information Science, 2008. RCIS 2008*, (Marrakech, Morocco), pp. 427–432, June 2008.

[30] U. M. Fayyad, G., and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[31] K. McGarry, "A survey of interestingness measures for knowledge discovery," *Knowledge Engineering Review*, vol. 20, pp. 39–61, March 2005.

[32] "Kdd." http://en.wikipedia.org/wiki/KDD, February 2009.

[33] L. Soibelman and H. Kim, "Data preparation process for construction knowledge generation through knowledge discovery in databases," *Journal of Computing in Civil Engineering*, vol. 16, pp. 39–48, January 2002.

[34] K. K. Loo, I. Tong, B. Kao, and D. Chenung, *Online Algorithms for Mining Interstream Associations from Large Sensor Networks*, vol. 3518 of *LNCS*. Springer Berlin / Heidelberg, May 2005.

[35] K. Römer, "Distributed mining of spatio-temporal event patterns in sensor networks," in *EAWMS/ DCOSS 2006*, June 2006.

[36] M. Halatchev and L. Gruenwald, "Estimating missing values in related sensor data streams," in *11th International Conference on Management of Data, COMAD 2005*, January 2005.

[37] R. Sterritt, "Discovering rules for fault management," in *Proceedings. Eighth Annual IEEE International Conference and Workshop on the Engineering of Computer Based Systems, ECBS 2001.*, pp. 190–196, April 2001.

[38] A. Boukerche, S. Samarah, and H. Harbi, "Knowledge discovery in wireless sensor networks for chronological patterns," in *33rd IEEE Conference on Local Computer Networks. LCN 2008*, (Montreal, Quebec, Canada), pp. 667–673, Oct. 2008.

[39] U. M. Fayyad, G. Piatetsky-Shapiro, and R. Uthurusamy, "Summary from the kdd-03 panel: data mining: the next 10 years," *ACM SIGKDD Explorations Newsletter archive*, vol. 5, no. 2, pp. 191–196, 2003.

[40] S. Pandey and P. Agrawal, "A survey on localization techniques for wireless networks," *Journal of the Chinese Institute of Engineers*, vol. 29, no. 7, pp. 1125–1148, 2006.

[41] M. Youssef, A. Noureldin, A. F. Yousif, and N. El-Sheimy, "Self-localization techniques for wireless sensor networks," in *IEEE/ION PLANS 2006, Position Location and Navigation Symposium*, (San Diego, California, USA), pp. 179–186, April 2006.

[42] A. Dunkels, F. Osterlind, N. Tsiftes, and Z. He, "Software-based online energy estimation for sensor nodes," in *Proceedings of the 4th Workshop on Embedded Networked Sensors, EmNets 2007*, (Cork, Ireland), pp. 28–32, 2007.

[43] G. Mathur, P. Desnoyers, D. Ganesan, and P. Shenoy, "Ultra-low power data storage for sensor networks," in *Proceeding of the Fifth IEEE/ACM Conference on Information Processing in Sensor Networks IPSN '06*, pp. 374–381, April 2006.

[44] Toshiba, "Toshiba 128-mbit (16*m*8*bits*/8*mx*16*bits*) cmos nand e2prom." http://www.datasheetcatalog.com/toshiba/67/, April 2009.