PhD Proposal

Spatio-Temporal Data Mining On Moving Objects In DBMS

Yahaya Bin Abd Rahim

A research proposal submitted in part fulfillment of the degree of

Doctorate of Philosophy in Geo-information

Promoter: Peter van Oosterom

Co-Promoter: Nanna Suryana Herman

Supervisor: Wilko Quak

GIS Section
OTB Research Institute for Housing, Urban and Mobility Studies
Delft University of Technology
November 15 2007

ABSTRACT

In this paper we present a pattern detection method for moving points object based on the spatio-temporal data types. However to proceed with our main goals, we need to develop the spatio-temporal data types and integrate these types of data in DBMS environment. Two processes are included in this research, there are integration spatio-temporal data types in DBMS environment and discover the patterns on moving objects in DBMS. We use bottom-up techniques to precede this research where we start with development of data types, integrate it and analyze it. In this research few basic techniques and models also are reviewed and discussed like spatial data, temporal data, data mining model and spatio-temporal data model. We also try to find the best techniques that can be adopted to improve the performance of the research especially in spatio-temporal data mining. In this research we also emphasize on the aspect of data mining in such a data model.

Keyword:
Spatio-temporal data types, moving point objects, spatio-temporal data mining, pattern

Chapter 1: INTRODUCTION

Time and space are important aspects of all real world phenomena. Database applications must be able to capture time and space varying in nature of the phenomena model. In conventional databases, attributes containing temporal or the spatial information are being manipulated solely by application programs, with little help from database management system. A spatio-temporal database is a type of database that supports aspects of both time and space. It offers spatial and temporal data types in its data model and query language.

Applications that rely on spatio-temporal databases can be found for examples, in geographical information systems, autonomous navigation, tracking, location based services and medical imaging.

Introduction and background in this chapter, drives this research towards "finding patterns for moving objects with spatio–temporal dataset in DBMS environment". Sub questions that support this research will be discussed in chapter 2 and chapter 3. In chapter 2 details state of the art technology in this research like spatial data mining, data mining, spatio-temporal data and DBMS will be further elaborated. Chapter 3 is the main focus of this research. We will discuss goal and methodology or techniques that will be used in this research. Whereas in Chapter 4, this research will touch on actions or activities that will be carried out.

1.1 Background

Spatial data in GIS is defined as elements that can be stored in a map, images, graph and tabular forms. Spatial data is required for sustainable development and optimized decision making. Considering that, types of data involved are complicated, time consuming and costly processes; therefore it is important to effectively manage subject of spatial data.

When temporal integrates with spatial, time variant and spaces are the factors that should be taken into consideration. This is because each time there is a movement; it is being recorded as a record that starts from the beginning until the end. The record became bigger as the movement became more frequent thus, giving impact on storage needed in storing the informations.

Efficient management of spatio-temporal data has gained much interest during these past few years [5,6,8,9] mainly due to rapid advancements in telecommunications (e.g., GPS, Cellular networks, etc.), which facilitates collection of large datasets such as information. Management analysis of moving object trajectories is challenging due to the vast amount of collected data and spatiotemporal. In many applications, the movements obey periodic patterns; i.e., the objects follow the same routes (approximately) over regular time intervals.

The problem of discovering periodic patterns from historical object movements is very challenging. Usually, the patterns are not explicitly specified, but have to be mined from the data. Patterns can be thought of as (possibly noncontiguous) sequences of object locations that reappear in the movement history periodically. Moreover, since we did not expect an object to visit *exactly* the same locations at every instant of each period, the patterns are not rigid but differ slightly from one occurrence to the next. The pattern occurrences may also be shifted in time (e.g., due to traffic delays or waking up late again). The approximate nature of patterns in the spatiotemporal domain increases the complexity of mining tasks. We need to discover, along with the patterns, a flexible description of how they vary in space and time.

1.2 Problem Statements

New concepts and methods are needed to extract more complete and detailed information from the vast repositories of spatio-temporal data that are accumulating. Methods for detecting spatial and temporal patterns across multiple data sets will enhance our ability to interpret spatial data and generate usable information, such as trends and causal relations.

Currently the efforts of spatial data mining are focus on either the spatial or the attribute of domain. Research accommodating both spatial and temporal data mining like moving objects is sparse only in methodology or data modeling not in implementation. Many efforts have been put forward in the study of spatial autocorrelation; the efforts specialize on the general patterns. A focus on spatial-temporal data mining could significantly expand the use of geospatial data in a variety of scientific areas, as well as in many practical applications such as weather prediction, transportation management, and environmental protection. In response to such a demand, this research plans on how to develop *an algorithm for finding patterns in moving point objects in a DBMS environment.* However the dataset that we are using for this research is in offline mode.

Since Database Management System (DBMS) is generic system and most of the organizations use DBMS as their information center because it provide value added to implement the pattern in DBMS environment such as flexibility, stability, few features and robust.

These are among the factors why the DBMS environment is chosen to implement the pattern. Pattern that was extracted from spatio-temporal should have capabilities with the options of DBMS especially the features of DBMS. The pattern also applied with the features of DBMS.

In this case we assume basic functionality handler like storage, indexing, update; query and etc are already implemented in PostgreSQL.

CHAPTER 2: STATE-OF-THE-ART TECHNOLOGIES

Output of this research will give a big impact on GIS application especially those researches that are related to spatial data. Pattern from spatio-temporal data is most important when processing a moving object data. The processing of moving object data will need to apply data mining techniques when extracting knowledge. In order to find the pattern, we need knowledge as an input to simulate the model. In this chapter, literature of spatial data mining techniques, associate rules, spatio-temporal data and DBMS are being introduced.

2.1 Data Mining

Data mining can be declared as discipline that concentrates on extensive database manipulation. As defined by various research papers, data mining is about process of searching hidden information that can be turned into knowledge thus could be used for strategic decision making or answering fundamental research question. It is most often used to help discover relationships, make choices, predictions and improve process. Data mining involves systematic analysis of data using automated method or computer algorithms in an effort to identify meaningful or otherwise interesting and potentially high-utility patterns, trends, or relationships in a data. Data mining draws on techniques from machine learning, database management and statistics to rapidly search for patterns in the data [15].

Data mining are most often used to help discover relationships, make choices, make predictions and improve processes. Data mining use algorithm to extract information and patterns derived by the KDD process [16]. Many interesting things can be found by using data mining that cannot be found by database queries such as “find people likely to buy my product”, “who are likely to respond to my promotion” etc. Data mining has very popular usage in discovering relationships problems. It started from well-known story about a large retailer who conducted a data mining experiment by looking at thousand of register receipts to discover which items people bought together. This analysis looks at concurrent events which sometimes is called “market-basket” analysis and sometimes referred as “link analysis”. Other than that, data mining can be used to make decision among available alternatives. For instance, data mining can be used to decide which customers will be the beneficiaries of allocated resources by applying classification technique to evaluate data availability and suggest or prioritize ‘best’ choices among the available alternatives. A prediction is a choice among alternatives available in future. History data can be used with data mining to make prediction. Finally data mining also can improve process to reveal aspects of business processes that are sub- optimal and estimate effects of proposed modification to these processes. In other words, data mining techniques can be used to learn the factors bearing on a decision and construct an application that uses those factors to help the management make those decisions in an objective, and consistent way.

In Data mining there are few techniques that can be used for mining. There are associate, clustering and classification. In associate model, it refers to analyzing the “market baskets”. It attempts to discover relationships or correlations in a set of items. Association models capture the co-occurrence of items or events in large volumes of customer transaction data. The apriori algorithm is the best solution if this technique is used to mine the data.

Clustering models are built for using optimization criteria that favor high intra-cluster and low inter-cluster similarity. This model can be used to assign cluster identifiers to data points. With classification models the data will be group into discrete classes and predicting which class the data belongs to. These data can be obtaining a model so that if the new data can be assign to the categories. This model is support by decision tree algorithm and Naïve Bayes algorithm.

2.2 Spatial Data Mining

When the data has relations with spatial data, the term becomes spatial data mining. In other words, spatial data mining is the application of data mining technique to spatial data. It will follow along the same functions in data mining; with the end objective is to find patterns in geography, meteorology etc. Figure 1 shows the data mining process.

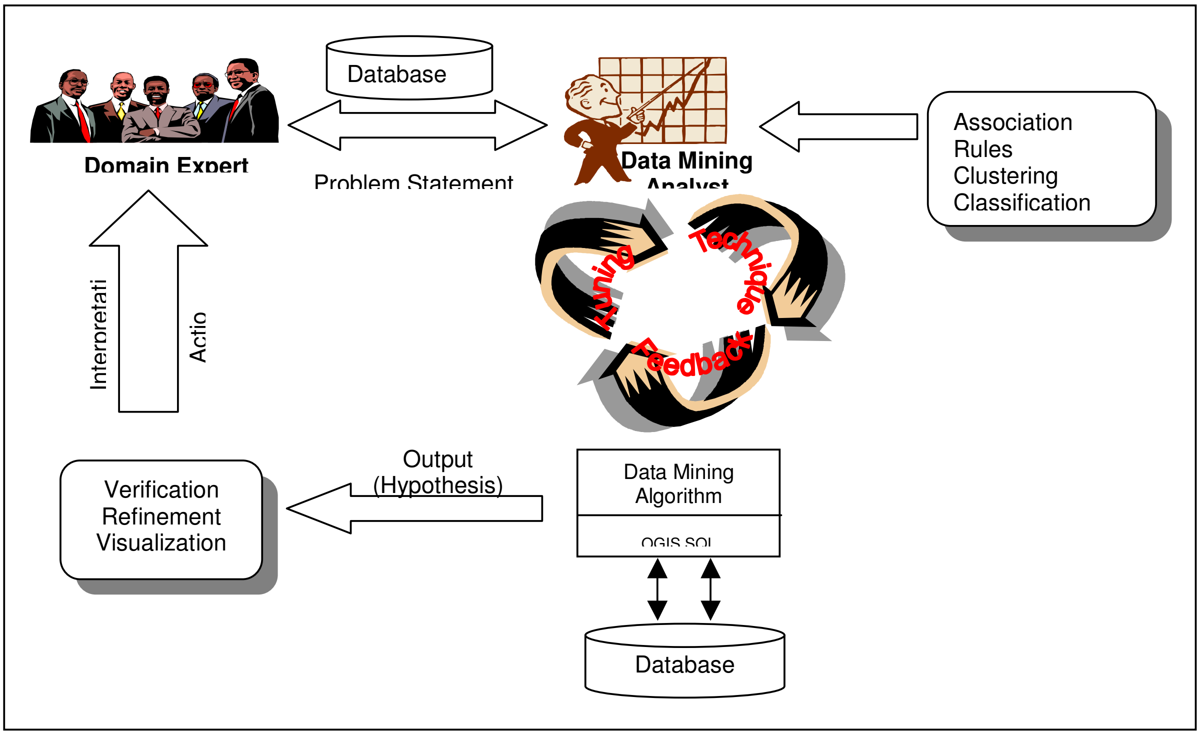

Figure 1: A Data Mining Process

Spatial data mining is the process of extracting implicit knowledge, spatial relations, or other patterns that are not explicitly stored in spatial databases [14]. Generally speaking, the spatial relations would not be stored in the database but concealed in multi-layer geographic data. To extract spatial description of the geometric features, some spatial operation must be performed which were involved in the process of spatial data mining. For instance, spatial data mining can be used in spatial (statistical) analysis to very large datasets such as finding cancer clusters to locate hazardous environment or to find new spatial and interesting patterns such as find locations that are unusual etc. Spatial patterns of interest here may include characterization of locations of a feature (e.g. crime) and its association with other spatial features (e.g. population density, distance to transportation network, etc).

As we know spatial database contains objects that are characterized by spatial location and/or extension as well as by several non-spatial attributes. The total number of spatial objects in the database sometimes can come up to a very large number. So the discovery process for spatial data becomes more complex than conventional data. Parallel with that extracting interesting pattern from traditional numeric and categorized data due to the complexity of spatial data types, spatial relationship and spatial autocorrelation. This applies to both the efficiency of algorithms and the complexity of possible patterns that can be found in a spatial database. One of the reasons is to mining in spatial database; spatial data mining algorithms have to consider the neighbours of some objects in order to extract useful knowledge.

2.3 Spatio – Temporal

A spatio-temporal database system manages data whose geometry changes over time. Applications that generate such type of data include surveillance applications, transportation systems, mobile communication systems and geographical and environmental systems, and so on.

Spatio-temporal datasets have some unique characteristics that make them different from traditional relational and transactional datasets. The difference is that changes can be continuous differently from traditional database systems where are assume the data changes through an explicit update [3]. To avoid continuous database updates on a spatio-temporal database, we need to store a description of the changes as function of time**.**

On the other hand, new concepts and methods are needed to extract more complete and detailed information from the vast repositories of spatio-temporal data that are accumulating. Methods for detecting spatial and temporal patterns across multiple data sets will enhance our ability to interpret spatial data and generate useable information, such as trends and causal relations.

Spatio-temporal data is usually modeled by extending temporal databases or spatial databases. That is, spatio-temporal data is modeled in two ways. First, we can add spatial properties and operations in temporal databases. The second way is to add temporal properties and operations in spatial databases.

Spatio-temporal data is generally used to describe complex objects. In most places, this data is stored as oriented-object database model and typically contains hierarchies of objects and includes concepts such as classes, inheritance, encapsulation and polymorphism. In spite of its flexibility in modeling entities, the resulting lack of uniformity may hinder data mining process. However when more complex data exists, potential approaches to data mining become unclear. The complexity of object-oriented databases makes the application of data mining techniques a challenging problem.

Compared to general data mining, spatio-temporal data mining has lots of special characteristics, such as rules that we can mine from it, similar patterns of change and spatio-temporal evolution patterns and so on. In addition, spatio-temporal data is usually large and complex, thus the data mining tasks we wish to perform are quite complex. It is important to discover efficient techniques that are suitable for spatio-temporal data. Sampling and biased sampling are two appropriate methods to improve the performance of data mining. Our spatio-temporal data model can describe both continuous and discrete change. Also, it is correspondingly easy for us to extract spatio-temporal knowledge from the model utilizing existing techniques.

For clearer figure 2 is illustrate the architecture of the spatio-temporal data-mining.

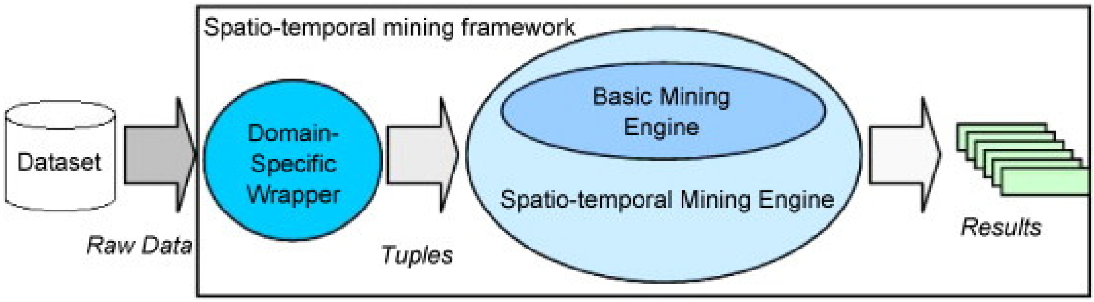

Fig. 2 of the spatio-temporal data-mining Architecture engine

2.4 Database Management System (DBMS)

A database management system (DBMS) is a collection of computer software or a complex set of software programs that are designed for controlling organization, storage and retrieval of data in purpose of managing databases. Typical examples of DBMSs include Oracle, DB2, Microsoft Access, Microsoft SQL Server, Postgres, MySQL and FileMaker. DBMSs are typically used by Database administrations in the creation of Database systems. Generally speaking, a DBMS facilitates the process of

- Defining a database; that is specifying the data types, structures, and constraints to be taken into account.
- Constructing the database; that is, storing the data itself into persistent storage.
- Manipulating the database like rename table, change field, indexing etc.
- Querying the database to retrieve specific data.
- Updating the database.

Figure 3 depicts a simplified database system environment. The illustrations show a DBMS acts as a mediator between user or application programs and the devices where data resides. DBMS software consists of two parts. The upper part processes the user query. The lower part allows one to access both the data itself and the metadata necessary to understand the definition and structure of the database.

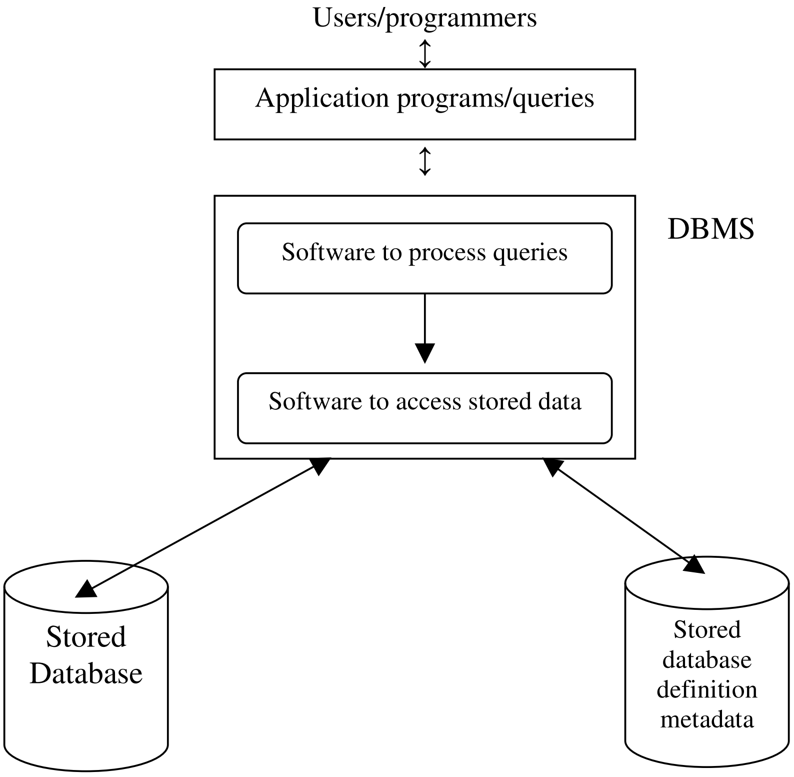
Users/programmers
Application programs/queries
DBMS
Software to process queries
Software to access stored data
Stored Database
Stored database definition metadata

Figure 3 A simplified database system environment.

Basically a DBMS includes a modeling language that defines the schema (relational model) of each database hosted in the DBMS. According to the DBMS data model most common organizations model could be the hierarchical model or network model or relational models but a database management system provide one, two or all three methods. Inverted lists and other methods are also used. The most suitable structure depends on the application and on the transaction rate and the number of inquiries that will be made.

The dominant model in use today is the ad hoc one embedded in SQL, a corruption of the relational model by violating several of its fundamental principles. Many DBMS's also support the Open Database Connectivity API that supports a standard way for programmers to access the DBMS.

A DBMS also include data structures (fields, records and files) optimized to deal with very large amounts of data stored on a permanent data storage device (which implies very slow access compared to volatile main memory).

Besides of that DBMS include a database query language and report writer to allow users to interactively interrogate the database, analyze and update it according to the users' privileges on data.

- It also controls the security of the database.
- Data security prevents unauthorized users from viewing or updating the database. Using passwords, users are allowed access to the entire database or subsets of it called *sub schemas*. For example, an employee database can contain all the data about an individual employee, but one group of users may be authorized to view only payroll data, while others are allowed access to only work history and medical data.
- If the DBMS provides a way to interactively enter and update the database, as well as interrogate it, this capability allows for managing personal databases. However, it may not leave an audit trail of actions or provide the kinds of controls necessary in a multi-user organization. These controls are only available when a set of application programs are customized for each data entry and updating function.

Since DBMS fully involve with the data, transactions is one of the most important data. A transaction ideally would guarantee the ACID properties, in order to ensure data integrity, despite concurrent user accesses, and faults.

- It also maintains the integrity of the data in the database.
- The DBMS can maintain the integrity of the database by not allowing more than one user to update the same record at the same time. The DBMS can help prevent duplicate records via unique index constraints; for example, no two customers with the same customer numbers can entered into the database. See ACID properties for more information.

DBMS accepts requests for data from application program and instructs operating system to transfer the appropriate data.

When DBMS is used, information system can be changed much more easily as the organization's information changes. New categories of data can be added to the database without disruption to the existing system.

Organizations may use one kind of DBMS for daily transaction processing and then move the detail onto another computer that uses another DBMS better suited for random inquiries and analysis. Overall systems design decisions are performed by data administrators and systems analysts. Detailed database design is performed by database administrators.

Database servers are specially designed computers that hold the actual databases and run only the DBMS and related software. Database servers are usually multiprocessor computers, with RAID disk arrays used for stable storage. Connected to one or more servers via a high-speed channel, hardware database accelerators are also used in large volume transaction processing environments.

DBMS's are found at the heart of most database applications. Sometimes DBMSs are built around a private multitasking kernel with built-in networking support although nowadays these functions are left to the operating system.

Features and Abilities of DBMS: One can characterize a DBMS as an "attribute management system" where attributes are small chunks of information that describe something. For example, "color" is an attribute of an object. The value of the attribute may be a color such as "red", "blue", "silver", etc. Lately databases have been modified to accept large or unstructured information as well, such as images and text documents. However, the main focus is still on descriptive attributes.

DBMS roll together frequently-needed services or features of attribute management. This allows one to get powerful functionality "out of the box" rather than program each from scratch or add and integrate them incrementally. Some features that include in DBMS:

- **Query ability**

Querying is the process of requesting attribute information from various perspectives and combinations of factors. Example: "How many 2-door taxis in Malaysia are green?"

- **Backup and replication**

Copies of attributes need to be made regularly in case primary disks or other equipment fails. DBMS usually provide utilities to facilitate the process of extracting and disseminating attribute sets. When data is replicated between database servers, so that the information remains consistent throughout the database system and users cannot tell or even know which server in the DBMS they are using, the system is said to exhibit replication transparency.

- **Rule enforcement**

Often one wants to apply rules to attributes so that the attributes are clean and reliable. Ideally such rules should be able to be added and removed as needed without significant data layout redesign.

- **Computation**

There are common computations requested on attributes such as counting, summing, averaging, sorting, grouping, cross-referencing, etc. Rather than have each computer application implement these from scratch, they can rely on the DBMS to supply such calculations.

- **Change and access logging**

Often one wants to know who accessed what attributes, what was changed, and when it was changed. Logging services allow this by keeping a record of access occurrences and changes.

- **Automated optimization**

If there are frequently occurring usage patterns or requests, some DBMS can adjust themselves to improve the speed of those interactions. In some cases the DBMS will merely provide tools to monitor performance, allowing a human expert to make the necessary adjustments after reviewing the statistics collected.

- **Meta-data repository**

Metadata or meta-data is information about information. For example, a listing that describes what attributes are allowed to be in data sets is called "meta-information".

2.5 Spatial Databases

A spatial database is a database that is optimized to store and query data related to objects in space, including points, lines and polygons [17]. While typical databases can understand various numeric and character types of data, additional functionality needs to be added for databases to process spatial data types. These are typically called geometry or feature. Spatial databases use a spatial index to speed up database operations [17].

Spatial databases architecture using a standard DBMS. There are layered architecture and dual architecture. In layered architecture the database is using the standard DBMS and in top of the databases there is spatial tools be a top layer of it. The figure 4 is the illustration of it.

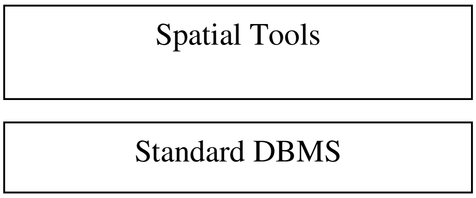

Figure 4 The Layered Architecture

In dual architecture in top layer is the integration layer that will integrate between standard DBMS with spatial subsystem in bottom layer. The figure 5 is the illustration of it.

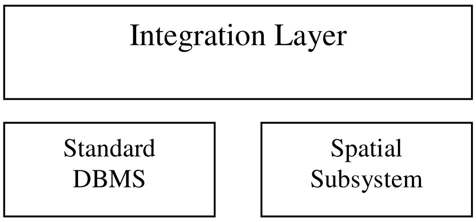

Figure 5 The Dual Architecture

The field of spatial databases has been an active area of research for over two decades. There are two common models of spatial information: field-based (e.g. raster) and object-based (e.g. vector). The field-based model treats spatial information such as altitude, rainfall and temperature as collection of spatial functions transforming space partition to an attribute domain. The object based model treats the information space as if it is populated by discrete, identifiable, spatially-referenced entities [17]. An implementation of a spatial data model in the context of Object-Relational databases consists of a set of spatial data types and the operations on those types.

Much work has been done on the design of spatial Abstract Data Types (ADTs) and their embedding in a query language, such as Spatial SQL. Spatial SQL has two parts: a query language to describe what information to retrieve and a presentation language to specify how to display query results. Users can issue standard SQL queries to retrieve non spatial data based on non spatial constraints. Moreover, they can issue Spatial SQL commands to inquire about situations involving spatial data and give instructions in the Graphical Presentation Language (GPL) to manipulate or examine the graphical presentation. The features of Spatial SQL have an Object Oriented flavor, such as the complex abstract data type spatial and its subtypes for different spatial dimensions. Paradise provides what can be loosely interpreted as an Object Relational data model. In addition to the standard attribute types such as integers, floats, strings and time. Paradise also provides a set of spatial data types including points, polygons, polylines, swiss-cheese and circles. The spatial data types provide a rich set of spatial operators that can be accessed from extended version of SQL Commercial examples of spatial database management include IBM DB Spatial Extender, Informix, Spatial DataBlade Module, Oracle's Universal Server with Spatial Option, Cartridge and ESRI's Spatial Data Engine [19].The functionalities provided by these systems include a set of spatial data types such as points, line segments and polygons and a set of spatial operations such as inside_ intersection and distance. The spatial types and operations may be made a part of a query language such as SQL, which allows spatial querying when combined with an ObjectRelational database management system. The performance enhancement provided by these systems includes a multidimensional spatial index and algorithms for spatial access methods: spatial range queries and spatial joins.

2.6 Pattern

If a work be done on a regular base and people know or expected the result we can call it a pattern. Most definitions of pattern state that pattern should be defined independently of scale but *Longman Dictionary* define pattern as the use of combination with time, sound, space and also to describe the structures behind our thinking. In software development, a pattern (or design pattern) is a written document that describes a general solution to a design problem that recurs repeatedly in many projects. Software designers adapt the pattern solution to their specific project. Patterns use a formal approach to describing a design problem, its proposed solution, and any other factors that might affect the problem or the solution. A successful pattern should have established itself as leading to a good solution in three previous projects or situations [nn].

In this research, we will look into the co-location pattern from the analysis of spatio-temporal data for moving points object. We try to find this co-location pattern because currently we cannot find any research regarding co-location pattern. The second issues why we choose this topics because we want to establish the propose framework with the mining process. If the frameworks can be accepted, the next actions are to declare few kinds of pattern with moving points object especially with the generic patterns. Other issues are to name the pattern, since there are few of definition about the pattern and no concrete ideas regarding how to name the pattern so it’s quite difficult for us to gives the name based on the analysis and this is the issues that we will raise as the limitation.

Chapter 3: PHD RESEARCH FOCUS

In order to smoothly and successful carry out the PhD research, this chapter summarizes and highlights the main research goal with some research question to be answered and also the methodology approach that will be used during this study.

3.1 Research Goal

Spatio-temporal data is the main component needed in order to find information especially with moving object or information that integrates with accumulative times. With spatio-temporal data as a source or dataset we need to extract to get the information and this extracting process is known as the mining process. The co-location pattern will be included in this extracting process. This new information from mining process will help us to minimize or solve the current problems in few application domains especially the domain that related with the changes over time.

The purpose of this research is to find the generic mining algorithms that can be use to produce patterns with moving object from spatio-temporal dataset. The dataset that we focus is this research for data type is in format vector point which in offline mode. This algorithm should be generic that can be applied in few cases like telecommunication, transportation and weather focus. Output of mining from this algorithm could be implemented inside a DBMS environment.

Since the dataset that we collect is not enough for us to extract the pattern and the stability of data log in GPS is not substance so we choose DBMS environment as the platform for mining the spatio-temporal data. We choose DBMS environment because DBMS is generic and stable and most companies in the world preferred DBMS environment in their information centre. The pattern that are produce should take into consideration the features in DBMS like query ability, rule enforcement, computation, automated optimization and the effectively data especially storage data. However in this research, we will only look at few of the features for example query ability, rule enforcement and computation features.

3.2 Research Questions

The research question of this PhD study is:

Main question:

- *"How do we find generic/moving patterns for moving point objects inside a DBMS environment"*

Sub questions:

- *"What is the technique or algorithm we can use to find patterns for moving point object?"*
- *"Why do we want to solve the problem inside DBMS environment?"*
- *"What are similarity case studies that we can use with this pattern?"*
- *"How we can apply these techniques efficiently inside a DBMS?"*
- *"What problem areas have a need for these techniques?"*
- *"What data model can we use to store moving object data?"*
- *"What is needed from the DBMS support?"*

3.3 Research Methodology

We use Hybrid methodologies which are combination within traditional method System Development Life Cycle (SDLC) or waterfall model with Rapid Prototype Model. The reason we use this methodology is because currently there are no specific statement that we can find in relation with moving point pattern with the spatio-temporal data set. Although similar research that were done for more than decades, most of them were not directly related with the spatio-temporal data sets. They use integration of data between spatial data with the temporal data to do analysis for telecommunication or other application.

In the first process, we use SDLC methodology to implement construction on the database or DBMS side. There are 5 phases that will take the action. Figure 6 shows the 5 phases in SDLC.

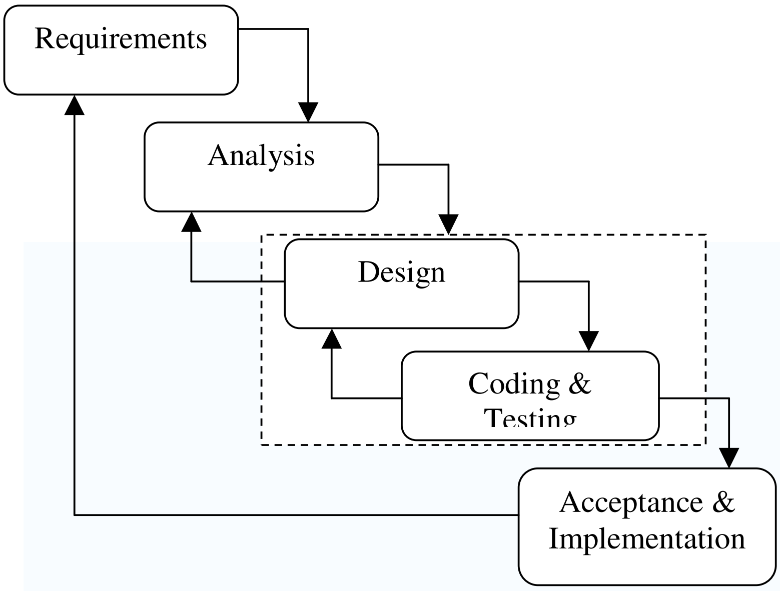

Figure 6: The illustration on 5 Phases in SDLC.

Phase 1 Preliminary Investigation

In this phase we try to define the problem definition. We need to find why we need to do this research and if it is in the market or already implemented in other application. In this phase we also need to clear the scope and objective of the research. By the end of this phase, we already have a clear picture of what we wanted to do and to be more specific we need to create research questions to help us find more details about our research.

Phase 2 System Analysis

After getting a clear picture on what we wanted to do then we analyze the present system by looking at specific documentation. The documents should contain information how the present system work and what it does. In relation to this, we are going to keep on doing literature review because until now we could not find any actual spatio-temporal data type for moving points object especially from GPS log files and because of that we need to do integration spatio-temporal data type in DBMS environment before we proceed with research on how to find the pattern.

Phase 3 System Design

Based on the report from phase 1 and phase 2, we proceed with the design on our research and system. We want to present how to find generic pattern in spatio-temporal data. This means new algorithm will be the core of the process. This algorithm will be implemented in mining spatio-temporal data.

Presently we study the join-less co-location mining algorithm, an apriori algorithm and associate rules for mining co-location patterns. Three methods will be used to study finding longer patterns; a bottom-up, level-wise technique and faster top-down approach [4].

In join-less co-location mining algorithm has three phases. The first phase converts an input spatial datasets into a set of disjoint star neighborhoods. The second phase gathers the star instances of candidate co-location from the star neighborhood set, and coarsely filters candidate co-locations by the prevalence value of the star instances. The third phase filters co-location instances from the star instances, and finds prevalent co-locations and generates co-location rules.

Apriori is designed to operate with databases which contain transactions. It is designed for finding associate rules in item sets. With association rule mining in the dataset, algorithm attempts to find subsets which are common to at least a minimum number of datasets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time and a step known as candidate generation, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a hash tree structure to count candidate item sets efficiently.

However we will use these two algorithms and associate rules method in the process of mining co-location patterns and integration spatio-temporal data types for moving points object in DBMS. We will find the best solution after we make few testing with different algorithms and method. Below are the framework design that we will use in development integration between spatio-temporal data sets from GPS log files in DBMS environment in figure 7a, and the framework for spatio-temporal mining process in figure 7b.

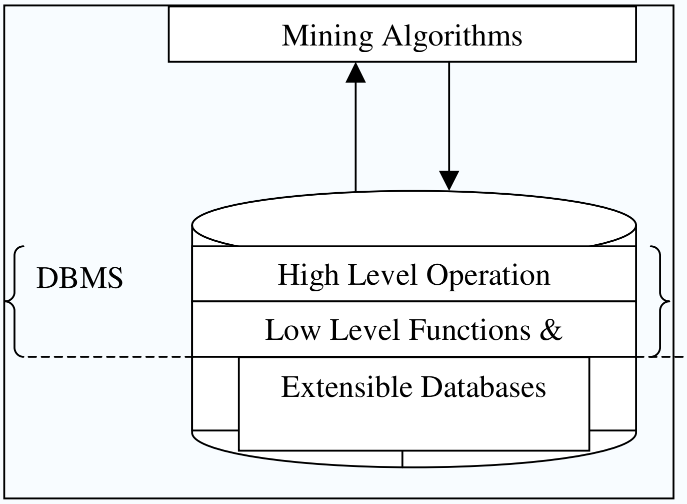

Figure 7a: A Framework for process Integration spatio-temporal data types in DBMS.

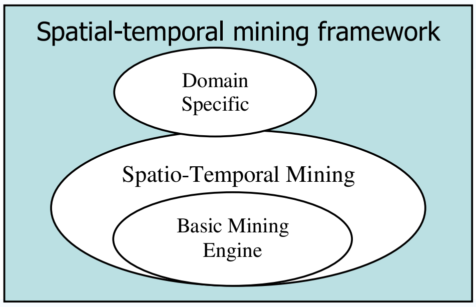

Figure 7b: A framework on Spatio-Temporal Mining Process

Phase 4 System Construction

In this phase, we will develop or do coding and testing. We would develop a prototype at this phase and test it with the data. In this research, two resources dataset will be collected and tested. The dataset comes from TWO cases. The first dataset data will come from other research and the second datasets will be collected during the case study. This two dataset will be mining based on the two methods either bottom-up or top-down approach.

For the first dataset, we will get existing research from other spatio-temporal data. This dataset will be analyze and extract to find pattern. The process could be done manually. After that, we were going to check the result and if the process did not follow our rules, its means there could be an error of logic error or coding error. If the error is logic error we will check with our design in phase 3 and change the design so that it follows our expected output; but if it is coding errors we have to find the bugs.

This process is Rapid Prototype where we develop and conduct test; if we need to change immediately we will change the prototype and if the outputs are what we expect the system can move to the last phase.

Phase 5 – System Acceptance and Implementation

In this phase, a system is implemented. All the dataset will be mined and the goal is to find generic algorithm in mining moving object with spatio-temporal data. The mining operator will be implementing in DBMS.

3.6 System Architecture

System architecture can be defined as a set of relationship between parts of a system. Many organizations define system architecture based on their own understanding. Therefore it can be concluded as relationship, integration or connection between hardware like components with software and concept. For example Carnegie Mellon University's Software Engineering Institute define system architecture as representation of a system in which there is a mapping of functionality onto hardware and software components, mapping of the software architecture onto the hardware architecture, and human interaction with these components. The National Center For Education Statistics define system architecture as description of design and contents of a computer system. If documented, it may include information such as a detailed inventory of current hardware, software and networking capabilities; a description of long-range plans and priorities for future purchases, and plans for upgrading and/or replacing dated equipment and software.

However systems architecture can best be thought of as representation of an existent or creation of a system, and the process or discipline for effectively implementing designs for such a system. The set of relations may be expressed in hardware, software or something else. A primary concerned with system architecture is the internal interfaces among the system's components or subsystems, and interface between the system and its external environment, especially the user.

For this research, system architecture will be based on integration of DBMS with spatial-temporal data and the framework of the spatio-temporal mining. Focus of the research will be based on bottom side first. We have to start from integration of DBMS with spatio-temporal data set from GPS log files. Reason why we have to start with bottom line because in the analysis phase, we currently found out there is none of spatio-temporal data type from GPS log files in real application. There are only data model on how to create spatio-temporal data types that integrated with DBMS. These means we will have to create the spatio-temporal data types from GPS log files first and then integrate the data types with DBMS.

After creating and integrating spatio-temporal data types with DBMS, we have to test the integrity of data and analyze the performance of data for examples test data on queries with Index or without Index key and test the data in bulk size. After completing the data integration and test the performance of data, we have to test the data with the function that were installed inside DBMS for example aggregation function and operational function. Due to the fact our objective model system is flexibility, we also propose to build ADT function and give user opportunities to build their own user defined function or aggregations. Below is the system architecture of this research.

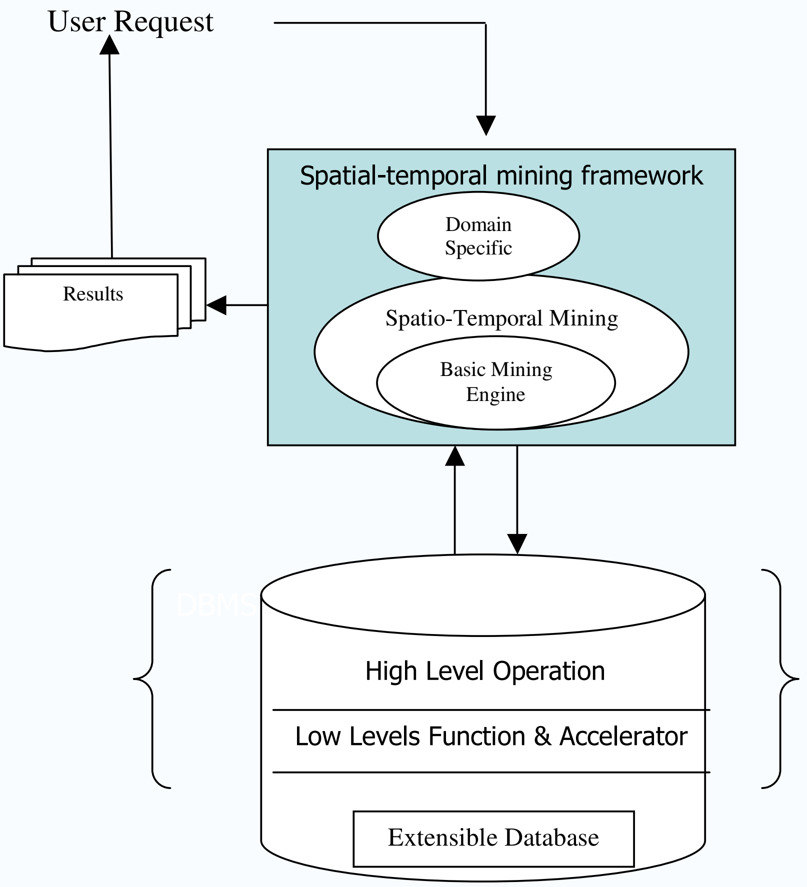

Figure 8: The System Architecture System Integration Spatio-Temporal Data with DBMS and Framework on Spatio-Temporal Mining Process

3.7 Research Scope and Limitation

Based on the architecture, scope of this research are the process of integration spatio-temporal data types in DBMS and producing the pattern by mining the spatio-temporal data in DBMS environment. Researcher has to find solution for integration spatio-temporal data types in DBMS environment as it is the basic steps in finding the pattern. Currently in the actual spatio-temporal data type it does not exsist and this is the reason why we need to do the integration before we could find solution in finding the pattern.

After completing the integration process, we could proceed with the second process in finding the pattern by analyzing moving point object data. GPS log files are the source files for spatio-temporal data types. However to be more reliable, we decide to find co-location pattern only with the moving point object data meaning the new data type is a point object with the time. If this research becomes successful, further research should look into more objects in spatial types and more than one pattern or the generic pattern.

3.8 Research Output

Based on the scope and limitation of research, we assume the research will produce few items as their output. New data type as spatio-temporal data are the first output and the first priority. After creating or producing new data type, we integrate DBMS environment and test with few existing functions in DBMS. We hope it would work like the normal databases. If it is successful, we will continue with other actions like building the spatial-temporal functions. The functions will helps user to identify pattern and existing database process such as aggregation, algebra and logic operation and etc. Defining pattern is the last output in this research and the output would be analyzed only for the co-location pattern. We will propose the framework for mining process and also propose new architecture for integration process.

CHAPTER 4: PhD RESEARCH PLAN

This is a rough idea how the PhD research will be carried out in a period of 4 years time. Unofficially this research will start on April 1st 2007. Like other common PhD plans, 60% of the total time will be spent on the PhD research work, and the rest of 40% will be concentrate on writing paper reports, attending conferences and education course and meeting with supervisor Wilko every week, with Prof. Peter and Prof. Nanna Suryana every month. The PhD Plan detail is as Figure 9.

Based on the Figure 9 below, writing PhD research plan will be complete by end of September. Followed with literature review and getting familiar with the tools such as database software, programming languages and gps. These tools would be used in research background studies and it will take about 3 months from September to November 2007.

January 2008 until Mac 2008, the research will concentrate in Data Understanding and write the summary of the literature review. Some revision needed for understanding the datasets model especially for spatio-temporal data sets. In data preparation, data collection will involve two types and it will be start from April until July. Reason why we need two dataset is because we try to adopt the pattern with two different cases. For the existing data in first dataset, we will look on the process to develop the spatio-temporal data types and to integrate the spatio-temporal data types in DBMS. In this case the intention is how the process is being done with the spatio-temporal data. For the second dataset we propose to collect data in Malaysia which are few possibilities sources will be input for this research like Celcom the Telecommunication Company, Remote Censor Department and Kuala Lumpur Town Council. Through the data, we are going to try to find the patterns of route during peak hours in Kuala Lumpur.

Perform the data mining and develop the algorithm will be starts from August until middle of November. After that data will test for processing, the output or solution from research will be present and at the same time the researcher will produce a paper for proceeding in conference or journal. The researcher will start to find at least one or few moving points patterns based on the output that we get from the processing data late stage. All the techniques will be test with the data from our previous case study. Each technique will be compared to each other in order to identify the best technique that can be applied. It will take about 7 month (January – July 2009). One paper will be produce at this stage and the comparison on techniques will be highlight. After choosing the best technique, a coding system will be implemented and we assume it will take around 5 months. This prototype will be tested and evaluate with 2 type of dataset. Model development will be implemented and tested.

In 2010, the implementation stage, system will be tested once again and modification will be made before deployment stage and implementing it in DBMS environment. This stage will take around 6 months to be completed. The last eight months of this research, will be spent for final preparation of writing the document, PhD Dissertation and preparation of defending ceremony. This research will be completed on April 2011.

<table>
<tr><th></th><th>Jan</th><th>Feb</th><th>Mac</th><th>Apr</th><th>May</th><th>Jun</th><th>Jul</th><th>Aug</th><th>Sep</th><th>Oct</th><th>Nov</th><th>Dec</th></tr>
<tr><td><b>2007</b></td><td colspan="3"></td><td>Unofficial PhD research</td><td colspan="7">Writing PhD Research Plan
&
Find Literature Review and also Familiarity with Research Tools(Software & Hardware)</td><td><b>Take Leaves</b></td></tr>
<tr><td><b>2008</b></td><td colspan="3">Write the Literature Review</td><td colspan="4">Data Collection (Propose in Malaysian)</td><td colspan="4">Perform Data Mining and Develop the Algorithm</td><td><b>Take Leaves</b></td></tr>
<tr><td><b>2009</b></td><td colspan="7">Select few techniques that will test based case study (Techniques studying and comparison)
& Prepare Paper Conference / Journal</td><td colspan="5">Code the technique with Programming Languages</td></tr>
<tr><td><b>2010</b></td><td colspan="4">Test and Evaluation of models, and make modification
&
Prototyping Plan</td><td colspan="4">Implementation & Test
In DBMS Environment</td><td colspan="4">Final preparation for writing the document
&
PhD Thesis Writing
&
Defense PhD Research</td></tr>
<tr><td><b>2011</b></td><td colspan="4">Final preparation for writing the document
&
PhD Thesis Writing &Defense PhD Research</td><td colspan="8"></td></tr>
</table>

1. Hui Yang and Srinivasan Parthasarathy. Mining Spatial and Spatio-temporal Patterns in Scientific Data. *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06)* - Volume 00:146, 2006.

2. Lubos Popelinsky and Jan Blatak. Toward mining of spatiotemporal maximal frequent patterns. In **Proceedings of ECML/PKDD Workshop on Mining Spatio-Temporal Data (MSTD).** Porto : UP, 2005. pp. 31-40. 2005, Porto.

3. Jia-Dong Ren and Jie Bao, Hui-Yu Huang. The Research On Spatio-Temporal Data Model And Related Data Mining. In *Proc. of International Conference on Machine Learning and Cybernetics*, 2-5 Nov 2003.

4. M.Nikos, Huiping Cao, G.George, H.Marios, Yufei Tao and W.-C. David. Mining, Indexing, and Querying Historical Spatiotemporal Data. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD'06:236 - 245, 2004.*

5. S. Ma and J. L. Hellerstein. Mining partially periodic event patterns with unknown periods. In *Proc. of 17th International Conference on Data Engineering, ICDE01*, pages 205–214, 2001.

6. B. O¨ zden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. In *Proc. of International Conference on Data Engineering*, pages 94–101, 1998.

7. X. Wei and H.-K. Huang. Research and Application of Spatio-temporal Data Mining Based on Ontology. *Proceedings of the First International Conference On Innovative Computing, Information and Control (ICICIC'06)*, Volume 2:535 - 538, 2006.

8. W.-C. Peng and M.-S. Chen. Developing data allocation schemes by incremental mining of user moving patterns in a mobile computing system. *IEEE Trans. Knowl. Data Eng.*, 15(1):70–85, 2003.

9. M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid. Periodicity detection in time series databases. *IEEE Trans. Knowl. Data Eng.*, 17(7):875–887, 2005.

10. Yeh T **S,** B de Cambray. Modeling Highly Variable Spatio-Temporal Data. 6* AustraliAsian Database conf. 221-230, 1995.

11. M.Erwig R.H, Guting M, Schneider, M. Vazirgiannis. Spatio-Temporal Data Types: *An Approach to Modeling and Querying Moving Objects in Databases*. GeoInformatica, 3(3): 269-296, Mar. 1999.

12. Peter Whigham. Spatio-temporal Modelling using Video Input. *Presented at SIRC 2000 – The 12th Annual Colloquium of the Spatial Information Research Centre, University of Otago, Dunedin, New Zealand,* December 10-13th 2000**.**

13. R.Gopalan, T.Nuruddin and Y.G. Sucahyo. Building a Data Mining Query Optimizer, *In Proceedings of the Australasian Data Mining Workshop, Canberra, Australia.* 2002.

14. R.Arbiol, Y.Zhang, V.Pala. Advanced Classification Techniques: A Review, Institut Cartografic de Catalunya (ICC),Parde de Montjuic, Barcelona, Spain, 2005.

15. C.Sanjay, S.Shashi, W.Wei Lu and T.Xinhong, Spatial Data Mining: An Emerging Tool for Policy Makers, An Interactive Research Grant from CURA, University of Minnesota, September 2000.

16. Usama Fayyad, G.Piatetsky-Shapiro,S.Padhraic, The KDD process for extracting useful knowledge from volumes data, Communication of the ACM, Volume 39,Pg 27 – 34, 1996.

17. S.Shekhar, S.Chawla, S.Ravada, F.Andrew, X.Liu,C.T.Lu.Spatial Databases-Accomplishments and Research Needs. *IEEE Trans. Knowl. Data Eng.*, Vol.11,No.1,pp. 45-55, January/February 1999.

18. Using wikupedia, http://en.wikipedia.org/wiki/Database

19. C.C.Xinmin, Data Models and Query Languages of Spatio_Temporal Information, A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Computer Science,2001.

20. P.Rigaux, M.Scholl, A.Voisard, Spatial Databases with Application To GIS. Morgan Kaufmann Publishers, San Francisco, 2002.