Case Study

Institution of MECHANICAL ENGINEERS

Journal of ENGINEERING FOR THE MARITIME ENVIRONMENT

Proc IMechE Part M:
J Engineering for the Maritime Environment
1–17

DOI: 10.1177/1475090214541426
pim.sagepub.com
SAGE

Online analysis process on Automatic Identification System data warehouse for application in vessel traffic service

Ming-Cheng Tsou

Abstract

The widespread application of the Automatic Identification System has had a revolutionary impact on navigation technology. In terms of its impact on a vessel traffic service, it provides rich and real-time data on the vessels, which can be used for identification, tracking, and monitoring of vessels. The Automatic Identification System rapidly accumulates large volumes of data, and these data contain a very large number of implicit maritime traffic rules and characteristics, and thus, effective methods are needed to discover the knowledge contained therein. In this research, data warehouse and online analysis process technologies, which are utilized by ordinary business entities for large-quantity business information analysis, are applied to analyze Automatic Identification System information collected through a vessel traffic service. Automatic Identification System raw data collected in a harbor area are used for analysis and post-processing using the geographic information system and database technology, and the processed data with time and space characteristics are stored in an Automatic Identification System data warehouse. In addition, online analysis process technology, the Geographic Information System, and pivot analysis are utilized to perform rapid multidimensional, meaningful high-level information inquiries, analysis of marine traffic characteristics, and rule discovery in marine traffic. These can be used as references for port development planning, traffic forecasting, navigation safety assessment, and making other policy decisions.

Keywords

marine traffic engineering, Automatic Identification System(AIS), Online Analysis Process(OLAP), data warehouse

Date received: 24 February 2014; accepted: 3 June 2014

Introduction

Since 2002, the International Maritime Organization (IMO) has requested that sea-going vessels over 300 GT and all passenger vessels be required to install the Automatic Identification System (AIS). Through very high-frequency (VHF) signals, AIS information is transmitted between vessels, from vessels to shore, or vice versa. It has greatly helped to promote ship navigation safety and prevent ocean pollution. The AIS can transmit accurate data every 3–10 s. AIS data include abundant dynamic ship information (e.g. receiving time, position, speed over ground (SOG), course over ground (COG), and rate of turn (ROT)), static ship information (e.g. Maritime Mobile Service Identity (MMSI) code, vessel name, nationality, ship type, ship length, tonnage, and destination), as well as attribute data that are space-related, time-related, and general; the content of the information is quite diverse. By analyzing the large volumes of AIS data of ship navigation, it is able to determine the obvious navigation rules of ships within the sea monitoring area. Therefore, the AIS has become increasingly important in vessel traffic service (VTS) marine traffic management.[1]

However, as the AIS contains a large amount of data and encapsulates many marine traffic characteristics, VTS now faces the same problems as general enterprises as described by American futurist Naisbitt,[2] who claimed, "We are drowning in information, but starving for knowledge." Without appropriate analysis theories and methods, it would be difficult to interpret and utilize the AIS data.

Department of Shipping Technology, National Kaohsiung Marine University, Kaohsiung, Taiwan

Corresponding author:
Ming-Cheng Tsou, 482, Chung-Chou 3rd Rd. Chi-Jin District Kaohsiung City, 80543, Taiwan.
Email: d86228006@yahoo.com.tw

For the purpose of this research, an AIS receiving station was set up near Keelung Harbor in northern Taiwan to model Keelung Harbor's VTS station and to receive ship AIS data from a similar sea monitoring area. This facilitates marine traffic engineering analysis and research processes and also serves as a case study. Data warehouse and online analysis process (OLAP) technologies, utilized by ordinary business entities for large-quantity business data analysis, are applied. Using the large amounts of raw AIS data received as a basis, through the Geographic Information System (GIS) and database technology, the temporal and spatial AIS data are subjected to an extraction, transformation, and load (ETL) process in order to establish an AIS marine traffic data warehouse as a data processing platform required for the subsequent analysis. It is also used to establish multidimensional, multilevel information inquiries with the operations of online analytical processing and pivot analysis that are necessary during the analysis. Finally, it is used to establish a display platform to generate meaningful high-level information that can quickly analyze the track itinerary distribution, traffic density distribution, traffic volume, velocity distribution, collision risk areas, and other marine traffic features of the water area near the port. These can provide references for port planning, traffic forecasting, navigation safety assessment, and making other policy decisions.

Previous work

With the proliferation of AIS, not only has marine traffic safety been enhanced but also a new and reliable approach to marine traffic survey has been found. This allows the VTS to have an extended monitoring range as well as greater functionality. It also enables better information gathering and processing, and the amount of data accumulated has increased rapidly. As a result, in recent years, researchers have begun to pay closer attention to the new application of AIS and the development of analysis technology in the management of VTS and have obtained significant results. In some of these studies, the analyses and utilization of AIS data were based on the location information, course, and speed of the dynamic data obtained from AIS, as well as the static data of the ship type, the ship's particulars, the calculated closest point of approach (CPA), the time to closest point of approach (TCPA), the ship course, speed of distribution, and rate of change; in other studies, the traffic flow or ship domain was constructed, and then the accident statistics were collocated in order to generate the collision probability model to assess and study the collision risk of ships. Mou et al.,[3] using the ship's size, speed, and course as the basis, utilized the linear regression model and the Safety Assessment Model for the Shipping and Offshore on the North Sea (SAMSON) model to conduct an analysis of the AIS data and subsequently evaluate collision risk in a harbor area. Qu et al.[4] proposed using three indices of ship collision risk, based on Lloyd's MIU AIS ship movement database, to quantitatively assess ship collision risks in the Singapore Strait: specifically, index of speed dispersion, degree of acceleration and deceleration, and number of ship domain overlaps. Silveria et al.[5] carried out an analysis based on the AIS data, characterized the marine traffic off Portugal, and conducted statistical analysis for the marine traffic within the Traffic Separation Scheme. Goerlandt and Kujala[6,7] analyzed the AIS data to obtain realistic input data for the traffic simulation: traffic routes, the number of vessels on each route, the ship departure times, main dimensions, and sailing speed. Montewka et al.[8,9] constructed the minimum distance to collision (MDTC) model according to information obtained from the AIS in order to assess the probability of ship–ship collision. Sormunen et al.,[10] based on the ship particulars in the AIS data, established a "spill" model that can model the penetration, spill probability, and size of the effects caused by a ship striking a chemical tanker and then explored the uncertainties inherent in risk analysis.[11] Ståhlberg et al.[12] considered a study of collision evasive action patterns in close encounters based on analysis of detailed ship movement data from the AIS system in order to gain insight into likely vessel speeds and course changes in close encounters as compared to actual encounter conditions. These parameters could be used as a proxy to check the credibility of the presented evasive maneuvering model, or to propose a more realistic model. Talavera et al.[13] proposed a novel method to quantify the uncertainty inherent in the paths that ships will navigate in the future using information provided by the AIS system on the paths followed by ships in the past. For management and analysis of mass data, some researchers have explored the regular patterns of maritime traffic. Aarsæther and Moan[14] applied computer vision techniques to automatically separate AIS data to obtain traffic statistics and prevailing features and also enable the production of a simplified ship traffic model. Zheng et al.[15] used two machine-learning methods, namely, clustering and graphical property analysis, to analyze characteristics of vessel traffic flow in the AIS database. Tsou[16] applied two data mining techniques, association rule mining and sequential pattern mining, for the analysis of AIS data.

With reference to the studies above, the specific features of this study are as follows: from the perspective of setting up the real-time decision support system, we apply the technology from the database management system and business intelligence to the management and analysis of mass data. It is very suitable for the Marine Traffic Observation network, which has long data collection periods and busy territorial waters, as it

can provide appropriate data processing, storage, and analysis. Moreover, since the associated risk assessment model is more complex in the setup stage, it is less convenient to use for decision-makers. The analysis proposed in this study can allow decision-makers to make up their own combinations of data dimension for analysis and is both simple and flexible. Since the analysis simultaneously considers the temporal data and the spatial data as well as the general attribute data, it is equipped with the multidimensional and multilevel features that can reveal characteristics not easily found in enormous amounts of data. In combination with the setup of the data warehouse, it can not only provide the required information for the OLAP itself but can also greatly simplify the convenience of information provision in combination with other assessment modes in the future.

AIS data warehouse and OLAP

Data warehouse technology overview

Marine traffic management authorities are facing issues such as faster traffic development, rapid increases in flow and cargo-handling capacity, a continuous increase in vessel size and speed, shipping management, and waterway construction. To deal with these issues, responding quickly to changes in marine traffic flow becomes the key to marine traffic management. It is also inadequate to only use the rapidly accumulating AIS data from a VTS for information searches and report generation in later days. Therefore, powerful data analysis tools are currently needed so that useful comprehensible information can be obtained from the data. This is so that informed decisions can be made using the data rather than based purely on intuition, and that scientific, rational decisions can be made in the face of rapidly changing marine traffic.

Currently, the AIS data in the VTS are only a traditional database and can hardly assist in upper management decision-making. The reason is that the traditional database only stores business process information at hand and lacks the support information needed for decision-making. In order to obtain decision-making support information, a data environment suited to decision-making needs to be constructed on the current database. We note that through the data warehouse, data from different systems can be integrated into a reliable, consistent, and continuously updated information collection. Through utilization of a relational database management system, traditional relational reports and queries can be applied to conduct historical data analysis, which directly supports a more complete multidimensional analysis. This is exactly what is needed to support flexible decision-making. Traditionally, a data warehouse has the following characteristics:[17]

1. *Subject-oriented.* The data in the data warehouse are organized by set subjects. The subject in this research refers to dynamic and static marine traffic data sent by vessels' AIS transponders to VTS stations.
2. *Integrated.* Based on extraction and sorting of distributed databases and then through system processing, integration, and arrangement, the data in the data warehouse are obtained. Data source inconsistency must be eliminated to ensure data in the data warehouse are consistent integrated information. In the future, through a VTS network, a more extensive and comprehensive marine traffic data warehouse can be constructed.
3. *Non-volatile*. The data in the data warehouse are mainly used for decision analysis. After data are in the AIS receiving database and are entered into the data warehouse, under normal circumstances, they are usually stored long-term. In other words, only a massive search operation is performed in the data warehouse, and there are few update and delete operations. In addition, the AIS receiving database needs to be loaded into the data warehouse periodically to perform updates.
4. *Time-variant*. Data in the data warehouse typically contain historical information. The system records overall marine traffic data and accumulated historical data on individual vessels from a certain point in the past until each current stage. Through these data, quantitative analysis may be performed and predictions may be made on marine traffic development history and future trends of the monitored waters.

Based on the above characteristics, we designed a proper ETL mechanism (or ETL rules) and stored the information obtained from the AIS database into the AIS data warehouse. The AIS data warehouse then serves as a structured data environment required for a decision support system and OLAP, in order to provide immediate decision-making support.

OLAP technology overview

With increasing amounts of accumulated data, every organization would like to reap from it the maximum business value (i.e. through retrieving useful information). To this end, the OLAP technology has been quickly adopted as it enables analysts to quickly, consistently, and interactively observe information from different perspectives, in order to interpret information in depth. This information is transformed from raw data and reflects actual business situations in a way that users can comprehend easily. The objective of OLAP is to analyze these data, search for model, trend, and exceptional situations and present different multidimensional views to users. It primarily provides decision-makers and upper management with information analysis and processing functionalities for a data warehouse, instead

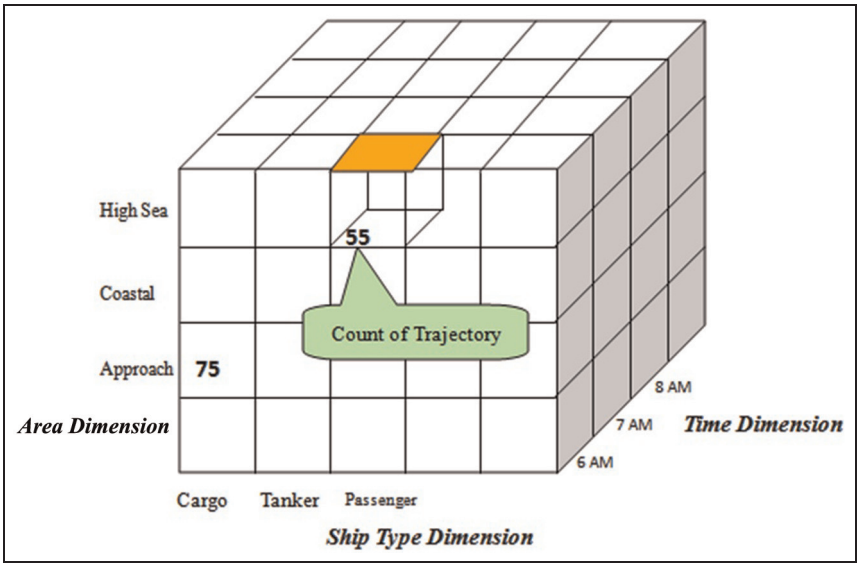

Figure 1. Marine traffic multidimensional data cube. It includes the dimensions of time, space, and ship type, and the measured number of vessels that appear.

of focusing on common routine business operations. Therefore, the relation between OLAP and the data warehouse may be viewed in the following way. The data warehouse is optimized for OLAP operations that include the aggregation or de-aggregation of information (i.e. "roll-up" and "drill-down," respectively) along a dimension of analysis, the selection of specific parts of a cube (i.e. "slicing" and "dicing"), and the reorientation of the multidimensional view of the data on the screen (i.e. "pivoting").[18]

The multidimensional data in the data warehouse utilized for OLAP can be presented in the form of a cube or hypercube based on its dimensions. Take a simple piece of AIS data as an example. Figure 1 shows its dimension table, fact table, and data cube. The fact table contains a large amount of basic facts that are of concern to VTS affairs. Dynamic data in the received AIS data can be regarded as the basis for the fact table. Ship type, time, and area represent three dimensions (they are each from dynamic data, static data, or voyage-related data). They are different points of view from which to analyze fact data, and they provide a numeric measure in the fact table as the basic information needed for the aggregation operation. An individual cube in the data cube may represent a specific numeric measure, such as the number of passed vessels, average navigation direction, or average navigation speed. Aggregation can be performed on these measures along each dimension for users' analysis. Accordingly, Figure 1 can be interpreted as the number of passenger ships that passed the high sea area at 6 a.m., which is 55, and number of cargo ships that passed the approach area at 6 a.m., which is 75. Through this structure and operation, traffic volume comparisons may be made in different combinations of time, area, and ship type. In addition, other dimension combinations and multiple operations (i.e. slicing, dicing, drill-down, roll-up, and pivoting) can be performed. This is to achieve multidimensional analysis, data ranking, prediction, evaluation, and decision support objectives. This also allows for subjective analysis verification and further enables the understanding of marine traffic volume distribution and changes in the monitored waters from multiple perspectives.

OLAP for the AIS data warehouse

This study is constructed based on the structure outlined in Figure 2. The temporal and spatial characteristics of AIS data were taken into consideration. The GIS, database management system, and data warehouse were utilized to process the data and to establish the AIS data warehouse. OLAP technology and the GIS were then used to perform analysis on the AIS data warehouse and to provide references for marine traffic management decision-making. We explain each of the important modules in this structure in the following sections.

Data collection

For this study, an AIS receiving station was set up about 1 km away from Keelung Harbor in northern Taiwan, as shown in Figure 3. It simulates Keelung Harbor's VTS station, in order to receive vessel AIS data within the monitored water area that is similar to Keelung Harbor VTS station. Through an RS-232 interface, the received AIS raw data were collected and saved in the database server, which served as a raw data storage platform. This allows for easy management and simplifies integration with other software. The data collection period was between 6 March 2013

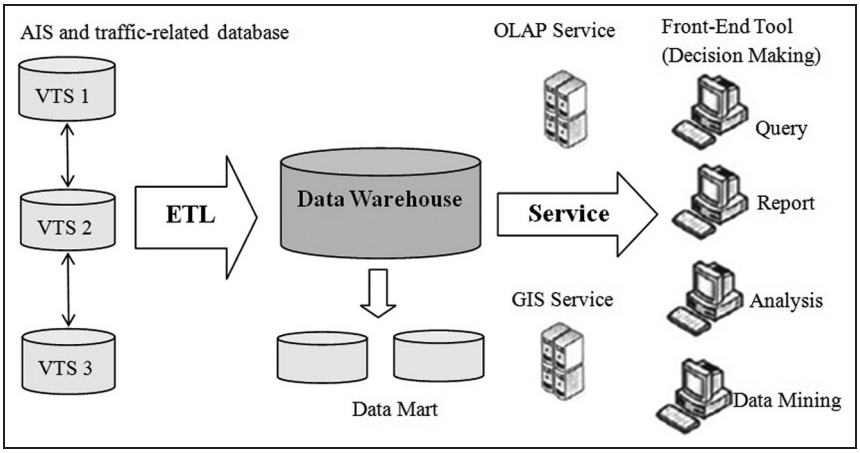

Figure 2. Diagram of the VTS marine traffic data warehouse infrastructure.
AIS: Automatic Identification System; VTS: vessel traffic service; ETL: extraction, transformation, and load; OLAP: online analysis process; GIS: Geographic Information System.

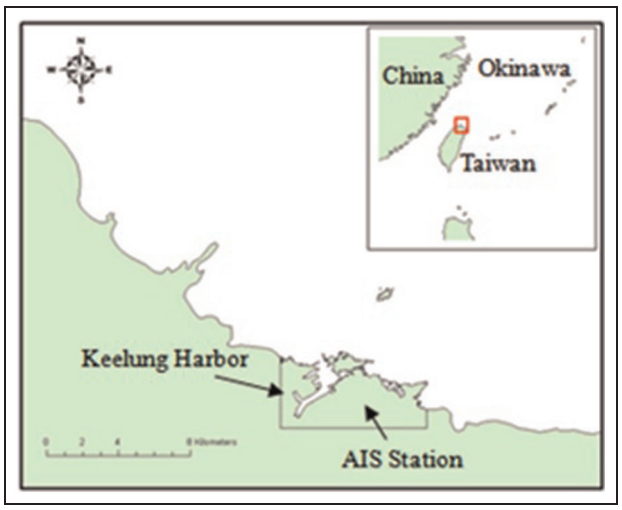

Figure 3. AIS station location map.
AIS: Automatic Identification System.

Figure 4. AIS received raw data.

and 28 May 2013, containing 3486 voyages and total static and dynamic messages up to 1.5 GB. The raw AIS dynamic data, as shown in Figure 4, contain a dense collection of independent data points. However, since much of the data were incorrect or had missing values, the first step was to perform data cleaning and filtering. In addition, using independent data points barely expresses the overall characteristics of a vessel trajectory. Owing to the substantial volume of data, certain important characteristics and rules of traffic flow cannot be shown. Therefore, it is necessary to perform trajectory reconstruction, in order to provide the basic data needed to perform OLAP on the AIS data.

Extraction, transformation, and loading of AIS data

Data extraction. The messages received from AIS transponders can be categorized into static information, dynamic information, and voyage-related information, which have different renewal rates in different time periods and under various navigation circumstances. Static information is entered when the AIS is installed and is only changed when the names of the ships are changed or when the types of the vessels have changed. Dynamic information is automatically updated through the sensor connected to the AIS. The voyage-related information is entered and updated manually during the voyage. In addition, the AIS can send safety-related messages that are not constrained by time. This research is based on Tsou's[16] research, and decoding is performed on the AIS message types 1–3 (target vessel position reports) and message type 5 (dynamic information and voyage-related information) messages.

Deletion of incorrect or missing values. Due to possible instrument failures or incorrect entry as a result of human error, there are often obvious inconsistencies in the AIS information. Upon processing, the following

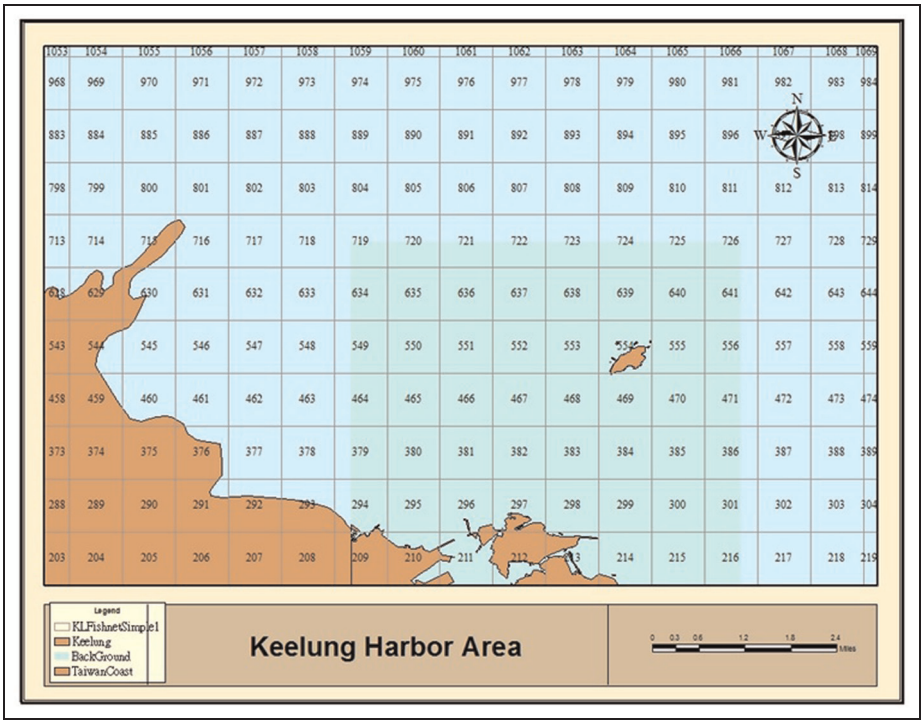

Figure 5. Grid division of the harbor area.

may be found: (1) unusual longitude and latitude numeric values inferring unreasonable locations (such as changes in latitude by more than 90°, in longitude by more than 180°, or incorrect datum inferring a vessel passing through land), (2) the final destination not being updated, (3) incorrect navigation direction (varying by more than 360°), or (4) incorrect navigation speed. These can all affect the statistical results. Since it is difficult for us to supply or correct the information, the incorrect information needs to be discovered and deleted to prevent mistakes in the analysis.

Establishment of a space relation. After decoding the AIS information, the spatial information is limited to latitude and longitude text records in the data table. The relationship between this information and the actual spatial location has not been established. The main objective of this step is to make a geographic grid of the harbor area and to establish the relationship between the AIS data and the harbor water area. Traditional marine traffic analysis mostly concerns pass line analysis and a less comprehensive analysis is conducted on the whole water area. This research uses the grid as an analysis unit and divides the studied water area into several geographic grids to facilitate the establishment of a spatial relationship with AIS information. We divide the area so that each grid has a size of $0.01° \times 0.01°$ (as shown in Figure 5) and each grid is assigned a unique number. Besides grid division, the water area is categorized into harbor, approach, and coastal areas based on different offshore distances and impacts on marine traffic to facilitate exploration of each area. Through GIS spatial processing, the dynamic data of each ship position are matched to a certain grid and a water area category to facilitate understanding of the areas passed in the trajectory and the spatial analysis that follows. This will further contribute to a holistic understanding of the studied water area.

Trajectory reconstruction. In the AIS application, the movement of a ship is often given by means of a finite set of dynamic messages, that is, time-stamped positions along with the MMSI. An important task is grouping and filtering these raw points arriving in the data stream in order to generate several meaningful *trajectories*, which are portions of the whole movement of a ship. Therefore, it is necessary to perform reconstruction on every vessel's trajectory (as shown in Figure 6). We define the following trajectory reconstruction processes as follows:

- *Individual vessel dynamic data sequential order*. Since the received AIS data stream is directly saved into the database, data for different ships appear in alternating order in the dynamic data table. Each individual ship's trajectory cannot be directly identified. Therefore, each ship's trajectory must first be separated, and the key to this is the ship's MMSI code and the time when the data are received. We use the following Structured Query Language (SQL) statement

$$\text{Select} * \text{from dynamic\_table order by MMSI, Receive\_Time} \quad (1)$$

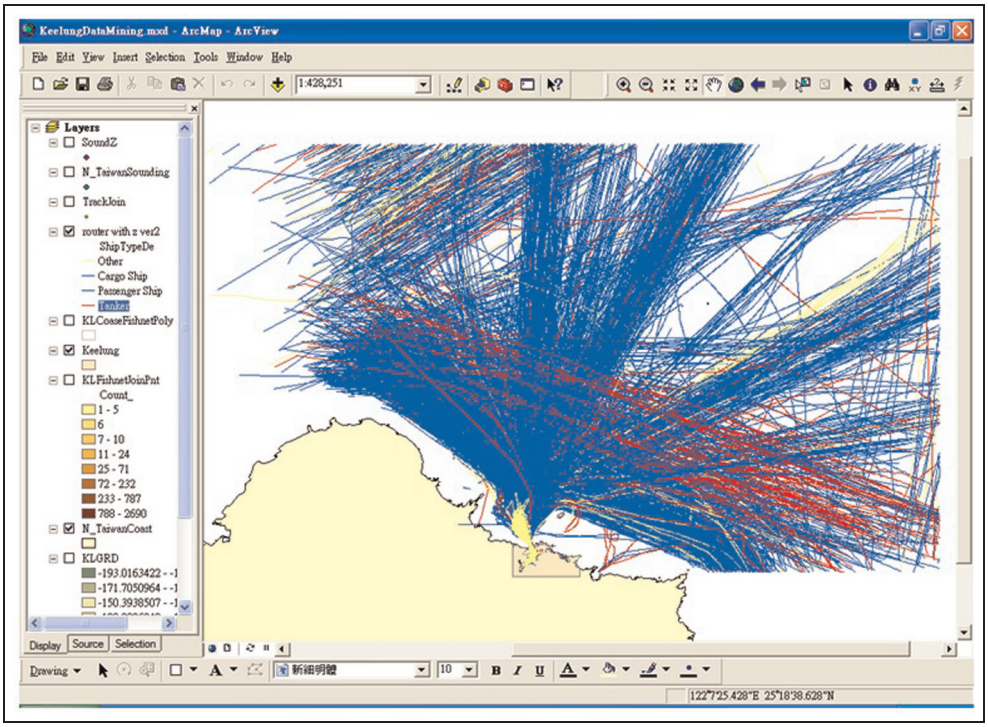

Figure 6. Ship trajectories obtained after trajectory reconstruction is performed on the dynamic data.

to process the data in the database and group the data of the same ship (by MMSI) and then order it sequentially based on the receiving time. After ordering the data in this way, it is still necessary to determine whether a reported time-stamped position must be considered as redundant and consequently discarded from the output trajectory. This research follows the suggestion of Goodwin[19] to filter out position points with ship speeds less than 3 knot, because these position points may be due to entering and leaving anchorage or picking and dropping pilots or other actions, instead of the usual navigation trajectories. Such filtering can increase the efficiency and accuracy of the analysis.

- *Temporal gap between trajectories (gaptime)*. At this point of processing, all historical data of the same ship in the dynamic data table are grouped together sequentially and may include several voyages or trajectories for entering and leaving the harbor. Therefore, it is necessary to establish the maximum allowed time interval between two consecutive time-stamped positions of the same trajectory for a single ship (by MMSI) to distinguish different trajectories. As such, any time-stamped position of the ship, received after a number of units more than the *gaptime* from its last recorded position, will cause a new trajectory of the same ship to be created.

Establishment of the AIS data warehouse

After performing the above-mentioned ETL processing, not only has unnecessary information been deleted but also useful information has been added. The data are now consistent or, in other words, clean. In order to allow OLAP to access the data and conduct analysis smoothly and to prevent additional data transformation work from lowering the efficiency or causing mistakes, we must transform the processed data into the data warehouse. The design of the system infrastructure is shown in Figure 2 and considers connections with other VTS stations and increases with traffic volume in the future.

Data source. The data source is the foundation of a data warehouse system. It may come from the AIS's receiving database or another marine traffic–related database. It may also come from the VTS's own station database or from an expansion to another VTS station database on a network in the future.

Data storage and management. The real key to the data warehouse is data storage and management. The data warehouse's organization management method differentiates it from traditional databases and determines its presentation format for external data. This research utilizes Microsoft SQL Server 2008 as the platform for data warehouse creation and data management and storage.

OLAP analyzer and GIS service. An OLAP analyzer is utilized to perform data integration on related attribute data within the main data. It is also organized based on a multidimensional model, to multi-perspective, multilayer analysis, and trend discovery. The Analysis

Service in the SQL Server Business Intelligence module is utilized in this research to create a multidimensional data cube. For the spatial data component of the main data, GIS software (ArcGIS) is utilized as the integration and analysis provision server.

Front-end tool. Front-end tools may include different kinds of reporting tools, search tools, data analysis tools, data mining tools, different data warehouses, and data mart–based application-software development tools. Microsoft Excel is utilized as the front-end tool in this research, and its pivot analysis table and pivot chart module are used. A Visual Basic for Application (VBA) program and the Component Object Model (COM) component of the ArcGIS Engine are used to connect with services provided by the GIS software (ArcGIS) to enable simultaneous processing of spatial data and attribute data, providing a search, presentation, and analysis operating platform. This feature also differentiates this research from an OLAP of a typical business problem.

Establishment of an OLAP multidimensional data model

The design of the data model is the core of data warehouse design. The multidimensional data model is the foundation of the data warehouse OLAP application.

Therefore, to establish the system, a conceptual model must first be designed to understand the system requirements clearly and to transform these requirements into an information structure. Then, a logical model is designed to transform the conceptual model into a data model that matches the actual system. One important principle of the design of a logical model of a data warehouse is to design a database structure that is easy to understand for easy search and analysis operations. In our logical model design, we tried to minimize the number of data tables and to simplify the relations between data tables in order to increase search efficiency. This simplified the database structure and also reduced the number of connections to necessary data tables during a search. One basic method of multidimensional data modeling is to utilize the star schema, which is composed of the fact table in the center and the dimension tables that surround the fact table. When using the star model to represent multidimensional data, despite the fact that the fact table contains a large amount of data with redundancy, the tables are reduced to the fact table and a limited number of dimension tables. The relations between them are simpler, making search operations easier. In consideration of the time and space complexity of marine traffic data, this research utilized the snowflake schema, which is evolved from the star schema and is more complicated in structure, as the modeling method for the logical model of the multidimensional data (as shown in Figure 7). The design method is as follows:

1. *Dimension table design*. The attribute values of the dimension table are typically text, discrete, and non-additive and will finally become the analysis and search criteria.
   - *Ship dimension table*. The data are mostly derived from the static data of the AIS and record information about the ship itself, of which ship length is a sub-dimension, in order to explore the distribution of ship lengths in the monitored water area. The MMSI in the ship dimension table serves as the primary key to establishing relations with the fact table.
   - *Voyage dimension table*. The data are mostly from the voyage-related data of the AIS, of which inbound/outbound (to represent whether the trajectory is an inbound or outbound ship) and draft are sub-dimensions. This is to locate the time and spatial distributions of the status and draft status of ships entering and leaving the port.
   - *Spatial dimension table*. As explained in section "Previous work," the dimension created using only numerical values of latitude and longitude is not suitable for processing space-related problems. Therefore, through GIS software, the numerical values of latitude and longitude are transformed into corresponding geographic cells to enable the creation of a spatial dimension through the geographic cells. We then designed different spatial granularities based on different classes. In addition, the water area is categorized into harbor, approach, and coastal areas based on various impacts on the marine traffic due to varying offshore distances (Table 1), in order to explore the differences in measured values of marine traffic in different spatial areas.
   - *Temporal dimension table*. The data warehouse stores historical data. As a result, the temporal dimension is the most basic dimension in OLAP. Considering the fact that the data collection period of this study is only 2 months, we included three levels in the temporal dimension design, which are day of the week (Sunday, Monday, …), part of the day (shown in Table 2), hour (1 a.m., 2 a.m., 3 a.m., …) in order to explore changes in the measured values of marine traffic under each time level in different periods.
   - *Ship-type dimension table*. As shown in Table 3, these data are included in order to understand the relationship between each ship type and the measured values of marine traffic in this sea area.
   - *Ship-length class dimension table*. As shown in Table 4, these data are included in order to analyze the relationship between each class of ship length and the measured values of marine traffic in this sea area.
   - *Draft-class dimension table*. These data are included in order to analyze the relationship between each class of draft ship and the measured values of marine traffic in this sea area.

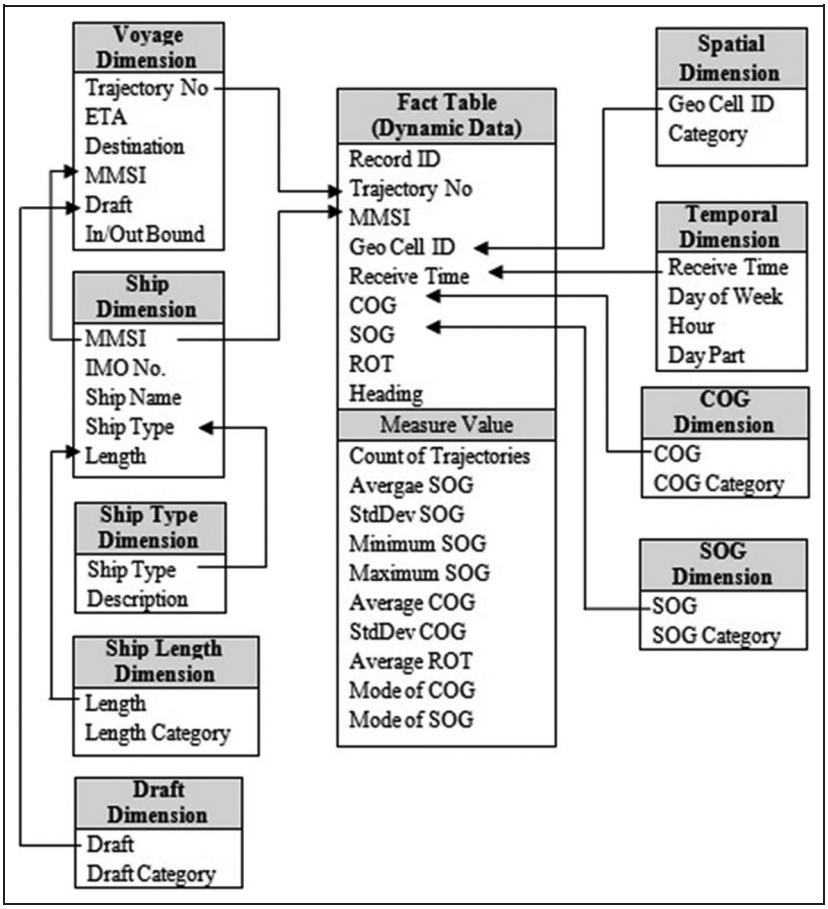

Figure 7. Marine traffic multidimensional data model.
MMSI: Maritime Mobile Service Identity; IMO: International Maritime Organization; COG: course over ground; SOG: speed over ground; ROT: rate of turn; ETA: estimated time of arrival.

Table 1. Geographic category.

Harbor
Approach
Coastal

Table 2. Day part.

0 a.m. to 6 a.m.
6 a.m. to 12 a.m.
12 a.m. to 6 p.m.
6 p.m. to 12 p.m.

Table 3. Type description.

Passenger (60–69)
Cargo ship (70–79)
Tanker (80–89)
High-speed craft (40)
Special craft (50)

- *COG-class dimension table*. As shown in Table 5, these data are included in order to understand the relationship between each category of COG direction and the measured values of marine traffic in this sea area.
- *SOG-class dimension table*. As shown in Table 6, these data are included in order to understand the relation between each category of SOG speed and the measured values of marine traffic in this sea area.

Table 4. Ship-length category.

Under 100 m
100–200 m
200–300 m
Over 300 m

Table 5. COG category.

N (337.5°–22.5°)
NE (22.5°–67.5°)
E (67.5°–112.5°)
SE (112.5°–157.5°)
S (157.5°–202.5°)
SW (202.5°–247.5°)
W (247.5°–292.5°)
NW (292.5°–337.5°)

Table 6. SOG category.

3–8 knot
8–13 knot
13–20 knot
Over 20 knot

1. *Design of the fact table*. The fact table is the most important table in the data warehouse structure. It directly reflects the topic of the data warehouse and includes the most important and fundamental information. For instance, this system analyzes marine traffic volume, is the cross point of each dimension table, and is the measure for a particular fact. It is the center of the star schema or snowflake schema and contains a large number of records. The columns contain not only key fields that connect with the dimension table but, more importantly, also measured value fields that correspond to the facts for recording. In this study, the main source for the fact table is dynamic data from received AIS data. The key fields used to connect with the dimension table include the following:
   - *Record ID*. This field is the primary key of the fact table used to identify individual ship location data. It is the foundation for ship trajectory construction and the main basis for performing OLAP drill-down analysis.
   - *Trajectory No*. Every record in the fact table can be matched with a trajectory and serves as a component for constructing the trajectory. The voyage dimension table of the received AIS data can be connected through this field.
   - *MMSI*. Every ship has a unique MMSI to identify which ship sends the position record. The ship dimension table and voyage dimension table of the received AIS data can be connected with this field, as the main reference for dynamic data construction.
   - *Graphic cell ID*. Every record in the fact table indicates ship navigation information at a certain position and time. Since the spatial information of the received AIS data is recorded separately in the latitude and longitude fields, numeric values in latitude and longitude fields are transformed into corresponding cells in the spatial region through a GIS software operation in order to connect with the spatial dimension table. This is an important field for spatial–temporal analysis on marine traffic information.
   - *Receive time*. This field records the receive time of each record in the fact table to connect with the temporal dimension table. It is an important field for spatial–temporal analysis on marine traffic information.
   - *COG*. This information is included in order to connect with the COG dimension table to facilitate COG state analysis of marine traffic.
   - *SOG*. This information is included in order to connect with the SOG dimension table to facilitate SOG state analysis of marine traffic.

The following numeric measure fields in the fact table serve as a basis for recording statistical information and analysis:

- *Count of trajectory*. This is used to record the number of ships or trajectory appearances under certain search conditions, such as the statistics of ship appearances in a certain period and a certain spatial region. This helps to determine the time period or spatial region in which the traffic is busiest. It is the main basis for traffic volume and traffic density analysis. Additional processing is required when performing an OLAP roll-up operation to avoid redundant calculations.
- *Average COG*. This is used to understand the ship's average COG under certain search conditions, such as the ship's average COG in a certain period and a certain spatial region. By doing so, we can discover the average flow direction of traffic flow and can provide flow direction references for traffic flow simulation. The average COG calculation differs from ordinary numeric data, and the built-in arithmetic mean method of ordinary OLAP tools cannot be applied directly to perform the calculation. For instance, when $COG_1 = 15°$ and $COG_2 = 345°$, their average COG does not equal 180° and should be 000° instead. We designed the following additional formulas to perform the Average COG calculation.

Since only the ship's course is calculated, the ship's speed is fixed. Assuming that there are n ships $(1 - n)$, and their navigation directions are $\theta_1, \theta_2, \ldots, \theta_n$, then the sum of the horizontal direction quantity X and the sum of the vertical direction quantity Y of all the navigation directions are calculated using the following formulae

$$X = \sum_{i=1}^{n} \sin \theta_i \times speed \tag{2}$$

$$Y = \sum_{i=1}^{n} \cos \theta_i \times speed \tag{3}$$

The average course angle (Cr) is calculated using

$$Cr = \tan^{-1}\left(\frac{X}{Y}\right) \tag{4}$$

The average COG (Avg_COG) is computed using

$$Avg\_COG = \begin{cases} Cr & (A \geqslant 0, \quad B \geqslant 0) \\ Cr + 360° & (A < 0, \quad B > 0) \\ Cr + 180° & (B < 0) \end{cases} \tag{5}$$

It is not very meaningful to use the average COG measured value directly. We cannot use it to determine the main direction of traffic flow, but we can use it to calculate the standard deviation (SD) of the COG and

determine the dispersion degree of the distribution of COGs.

- *Average SOG*. This is used to understand the ship's average SOG under certain search conditions, such as the ship's average SOG in a certain period and a certain spatial region. This enables the understanding of traffic flow speed in a certain period and in a certain region and may provide flow speed reference data for traffic flow simulation.
- *SD of SOG*. This is used to understand the dispersion degree and complexity degree of the distribution of SOG and may provide flow speed reference data for traffic flow simulation.
- *SD of COG*. This is used to understand the dispersion degree and complexity degree of the distribution of COGs. Using the SD of COG, the traffic flow encounter state in the area may be determined. For instance, when the SD is near 0°, the traffic encounter state of the area is mostly in the same direction. If the SD of SOG is also utilized, then whether or not it is mostly of an overtaking state can be determined. If the SD is between 30° and 70°, then the traffic flow state in the area would be mostly crossing or conversion. If the SD is more than 70°, then the traffic flow encounter state can be considered as close to head on. Through the use of the SD of COG along with the traffic density value, we can understand the traffic flow encounter state and then discover hot spots that have more complex traffic and are more likely to be encountered.
- *Minimum/maximum SOG*. This is used to understand the maximum/minimum SOG distribution and to provide flow speed reference data for traffic flow simulation.
- *Mode of COG/SOG*. Through appropriate classification, the mode of the related measured value of each dimension can be found. This can assist in understanding the major flow direction and speed of traffic flow in the area or under certain search conditions.
- *Average ROT*. This is used to understand the distribution of average ROT and can provide flow speed reference data for traffic flow simulation.

Demonstration and explanation of results

The major difference between the AIS data warehouse and those of other enterprises is the inclusion of temporal and spatial features. Ordinary business data warehouse software and OLAP tools typically only take attribute data into consideration and do not take spatial data into consideration, and thus fail to highlight the information implied by the AIS data in the spatial dimension. In order to highlight the spatial information, the GIS software needs to be relied on for processing, analysis, and final presentation. Therefore, the role of GIS in this research is not only as a spatial data warehouse but also in spatial processing, providing an analysis service, and serving as a spatial dimension information presentation platform. The integration of the GIS and data warehouse systems may be described as this research's most significant feature. The major analysis methods are pivoting and spatial visualization. Decision-makers do not need to consult information technology personnel and can conduct quick search and analysis procedures on different combinations of dimension interest on their own. This allows them to uncover not only regular patterns but also anomalous values. An arbitrary pass line is also possible for the observation profile. Since all the data in the data warehouse have been "cleaned," this analysis can be completely performed in almost real time. There are many other possible combinations, but they cannot all be explained owing to the limitation of pages in this article. Several other representative multidimensional searches are explained below.

Spatial and temporal dimensions and measured values of traffic volume (count of trajectory) analysis

The calculation and presentation of spatial dimensions in OLAP cannot be conducted solely using regular tools. If only the pivot analysis table is used to present search results, the spatial dimension component cannot be represented in a more meaningful way. As shown in Figure 8, a regular pivot table is used to construct the measured value of traffic volume (count of trajectory) through the spatial dimension of coastal area and the temporal dimension. Therefore, it is necessary to rely

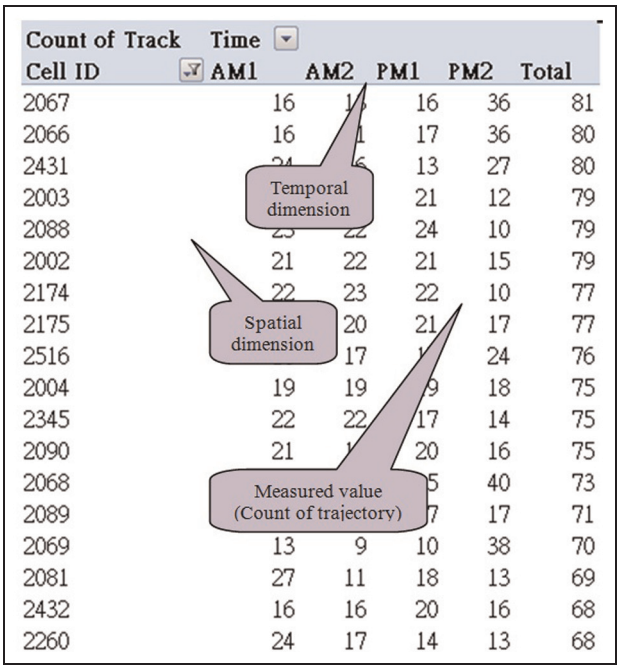

Figure 8. Spatial and temporal dimensions and measured values shown in the pivot table.

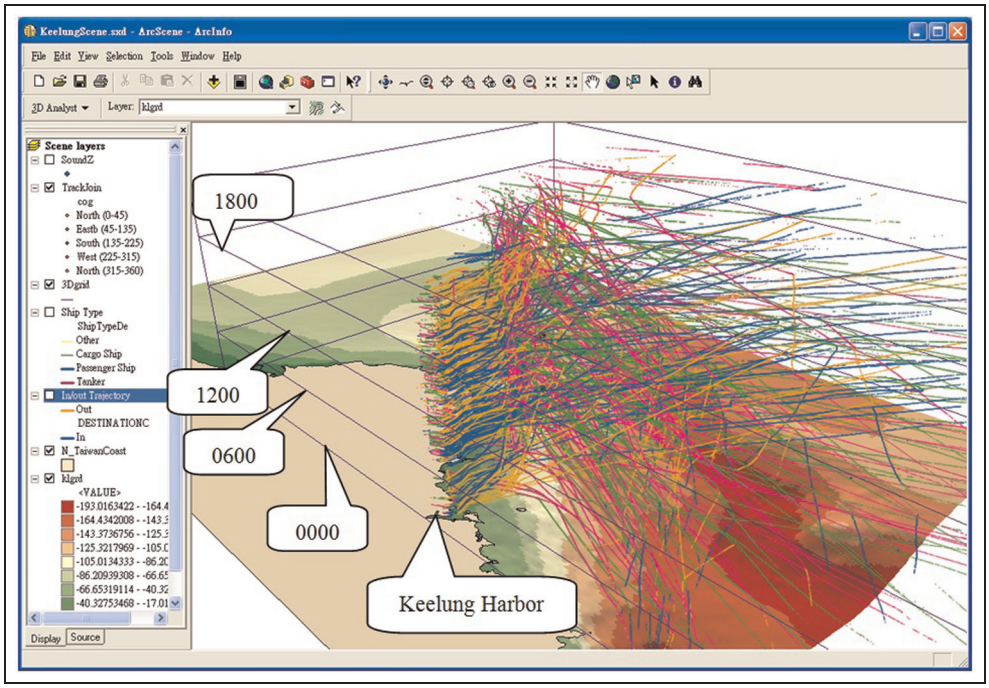

Figure 9. Presentation of GIS spatial and temporal dimension information with pivot table integration.

on GIS software to perform further processing and to transform the attribute value that has a spatial connotation into a specific spatial position. However, regular GIS software is less capable of processing the temporal dimension. We intend to conduct simultaneous analysis on the time and space dimensions, which is hard to achieve with current analysis software. In this research, through VBA and the ArcGIS Engine COM component, an Excel pivot analysis table is integrated with ArcScene three-dimensional (3D) analysis and presentation functionalities. As shown in Figure 9, the temporal data are matched with the elevation value of the 3D data. As a result, data at dawn are matched to a lower elevation, while data in the evening are matched to a higher elevation. In addition, a 3D flying method is simulated to perform roll-up and drill-down data analysis to better understand the spatial and temporal distributions and historical changes in the marine traffic volume, such as when and where there is greater traffic volume and when traffic volume is more complicated, as well as the ship-type distribution or port entry and exit states.

Combination and analysis of spatial, temporal, and other dimensions

Analysis related to spatial and temporal dimensions cannot be fully presented solely through a pivot analysis table and needs to be presented with the assistance of a GIS.

1. Relationship between traffic volume and spatial and temporal dimensions and port entry and exit dimensions. As shown in Figure 10, if the gate line of the port is regarded as the observation profile, of which the blue track represents ships entering the port while the others represent ships not entering the port, then we can find that ships enter the port mostly between 7 and 9 a.m. in the morning, while ships exit the port mostly between 4 and 6 p.m. in the afternoon.
2. Relationship between traffic volume and spatial, temporal, and ship-type dimensions. As shown in Figure 11, the red track represents a tanker, the green track represents a cargo ship, and the yellow track represents other types of ships. In this figure, we find passenger ships leave the port mostly between 10 and 11 p.m. at night and navigated in a northwestern direction, while tankers usually enter the port after anchoring (the vertical purple line), and cargo ships are distributed more evenly.
3. Relationship between traffic volume and spatial, temporal, and COG dimensions. Figure 12 shows the COG distribution of every track. The blue line represents the port entry track (south bound), and the yellow line represents port exit track (north bound). The traffic separation scheme outside the port clearly achieves the traffic diversion effect.
4. Relationship between measured COG values and spatial and temporal dimensions. The main focus of research here is the SD of COG. Through this numeric value, the navigation encounter state of ships in the water area can be determined as a basis for risk assessment. As shown in Figure 13, the darker area represents traffic flow encounters that are mostly in a head-on state, the lighter area represents those in crossing state, and the lightest area

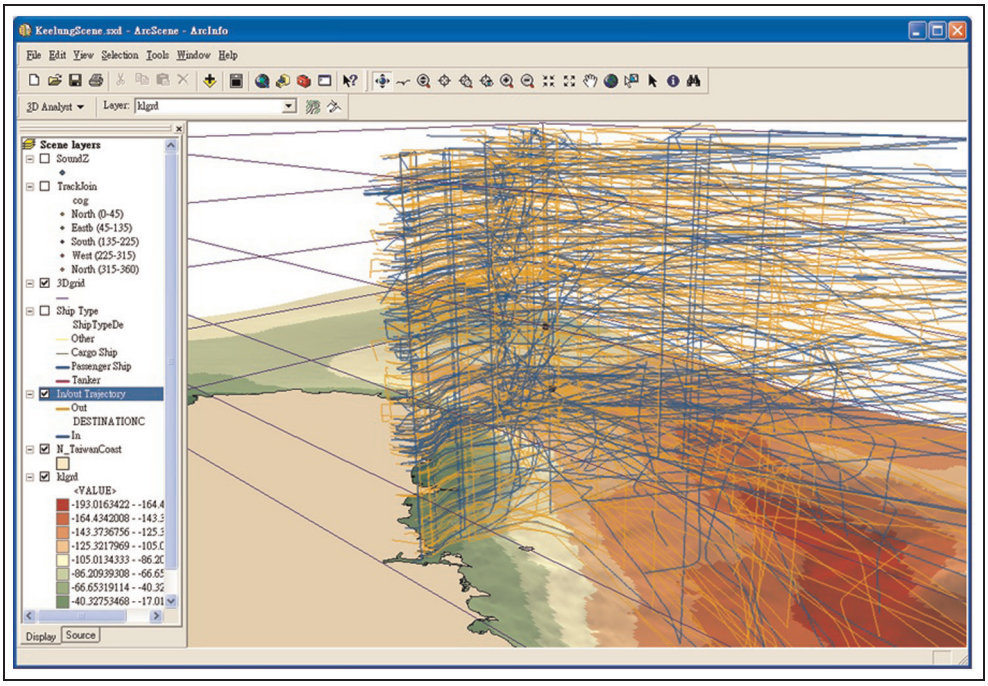

Figure 10. Spatial and temporal dimensions and inbound/outbound ships.

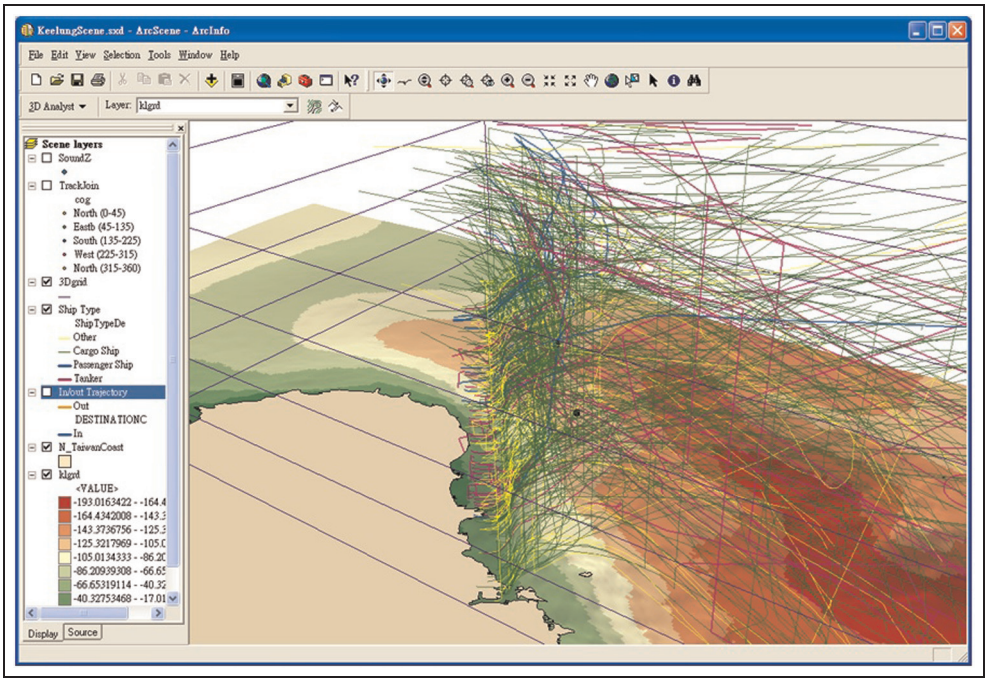

Figure 11. Spatial and temporal dimensions and ship types.

represents traffic flow in the same direction. A same-direction traffic flow evident in Area 1 can be identified in the picture. If it is compared with the inbound flow in Figure 12, then we can observe that this area belongs to part of the port entry traffic flow.

The encounter state of traffic flow in Area 2 is of the crossing type, while those in Areas 3 and 4 have more head-on encounter states. As the two places also have higher traffic densities, this serves as a basis for consideration of creation of a new traffic separation scheme in the future.

5. Spatial, temporal, and SOG measured values. The measured values here may include the maximum, minimum, average, SD, and mode of SOG to explore the characteristics of the distribution of

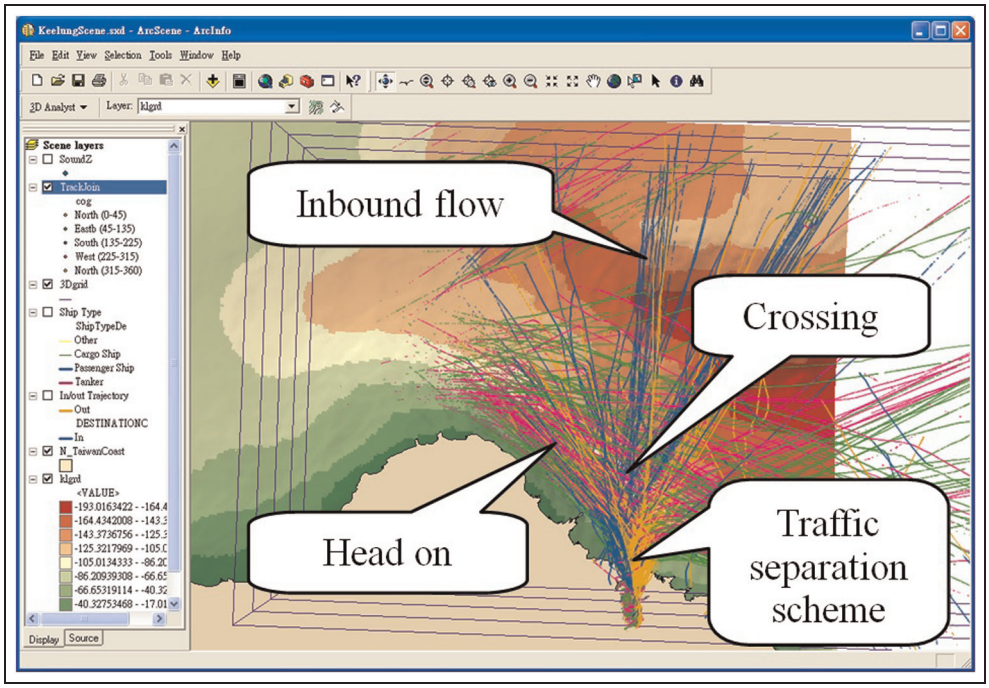

Figure 12. Spatial, temporal, and COG dimensions.

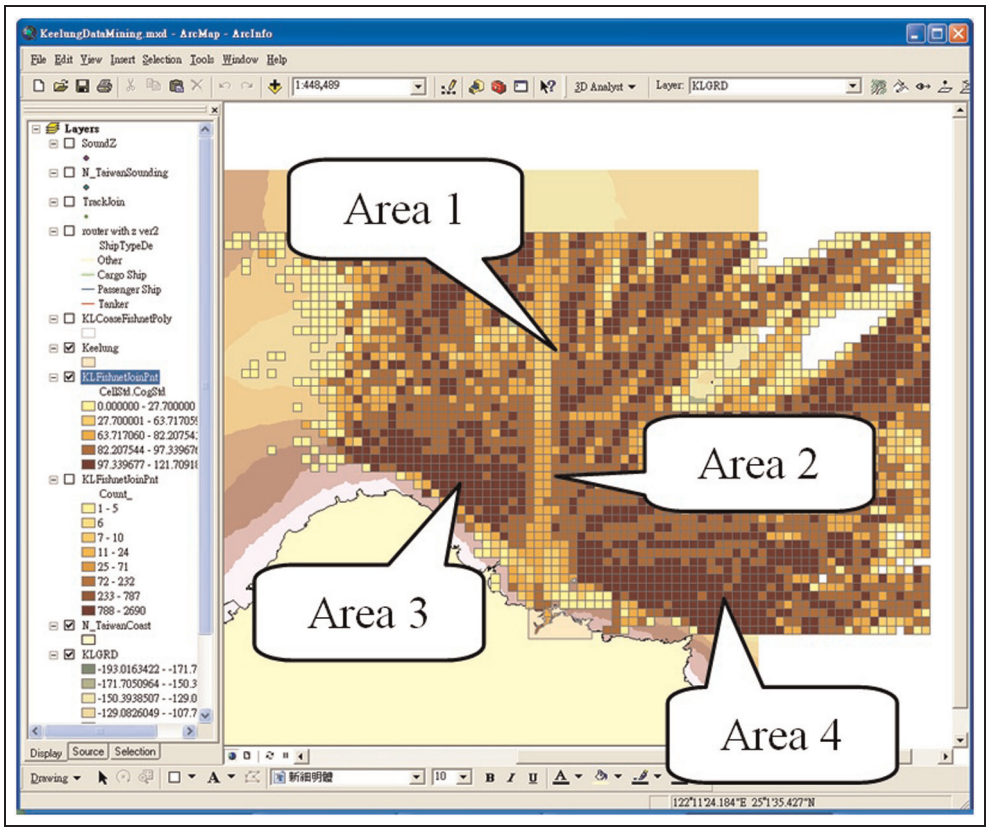

Figure 13. Standard deviation of COG.

traffic flow speeds. As shown in Figure 14, the darker color represents higher SOG measured values, while the lighter color represents smaller SOG measured values.

Other non-spatial dimension pivot analysis

This part belongs to traditional OLAP. We mainly perform the search using a pivot analysis table along with a pivot analysis diagram for presentation.

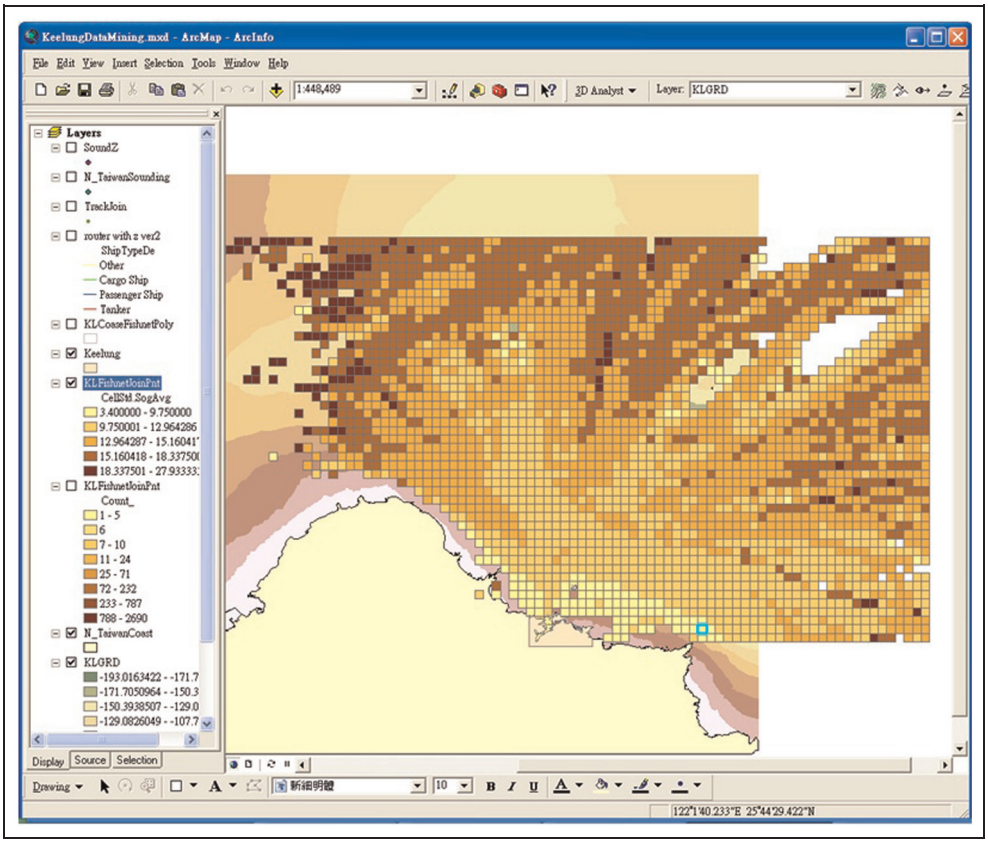

Figure 14. SOG measured values.

1. Traffic volume analysis of the day of week, ship type, and port entry and exit dimensions. As shown in Figure 15, the port entry traffic volume is higher on Sunday and Wednesday, while it is higher on Wednesday for passenger ships and higher on Tuesday for tankers.
2. Traffic flow analysis of draft dimension and port entry and exit dimensions. As shown in Figure 16, ship draft is mainly between 5 and 10 m (accounting for 77.6% of ships). The others account for small ships with drafts of 5 m or less.
3. Traffic volume analysis of the ship length, ship type, and port entry and exit dimensions. As shown in Figure 17, regardless of the ship type, ships entering the port mostly have lengths between 100 and 200 m (accounting for 61.2%), and few ships have lengths exceeding 300 m. Through this approach, extreme or anomaly values can be filtered.
4. Analysis of different levels of time periods. By establishing different levels of time periods, we can conduct searches based on different granularities of time. As shown in Figure 18, in level 1, we divide a day into four sections, and in level 2, a day is further divided into 24 sections. In doing so, abstract high-level (part of the day) and detailed low-level (hourly) analysis can be achieved to conduct roll-up and drill-down operations.

Conclusion

The accurate and detailed nature of information from an AIS provides abundant reference data to the VTS for information analysis and local assessment on ships navigating in the water area. As such, management authorities should no longer rely on experience and instinct to make decisions but, instead, they should utilize information tools to obtain accurate information from the data for decision-making and thereby enhance the resultant decision quality. In this research, data warehouse and OLAP technologies utilized by typical enterprises for mass business data analysis are applied to AIS data collected through the VTS. Unlike regular processing approaches adopted by typical businesses, this research specifically utilizes GIS and database technologies to process spatial- and temporal-related data in the AIS database. It allows marine traffic data in the sea area to be analyzed, and fast, multidimensional analysis searches to be conducted on spatial and temporal data through OLAP technology and the pivot analysis table. This consequently provides the necessary platform for data mining and other analytical models to obtain marine traffic regularity and trend information, providing a high-level information service to the VTS.

Currently, the international VTS is moving in the direction of a Vessel Traffic Management and Information System (VTMIS). The VTMIS not only

InOut In Bound

%	Type				
Day	Cargo Ship	Passenger Ship	Tanker	Others	Total
Sun.	14.89%	0.80%	0.54%	1.69%	17.92%
Mon.	13.24%	0.46%	0.39%	1.04%	15.14%
Tue.	11.51%	0.44%	0.85%	1.88%	14.67%
Wed.	14.29%	1.06%	0.39%	1.46%	17.21%
Thu.	7.26%	0.36%	0.62%	1.38%	9.62%
Fri.	11.39%	0.66%	0.43%	1.10%	13.58%
Sat.	9.31%	0.61%	0.70%	1.24%	11.86%
Total	81.89%	4.39%	3.92%	9.80%	100.00%

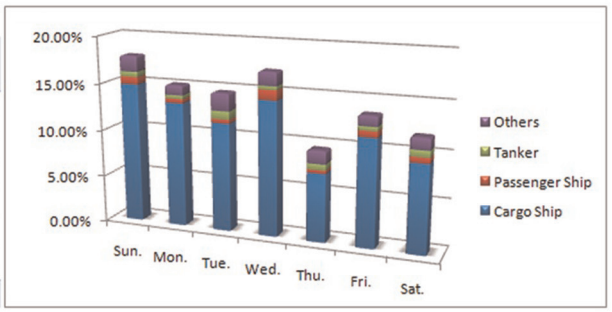

Figure 15. Temporal dimension, ship type, and inbound/outbound pivot table.

%	I/O		
Draft	In Bound	Non In Bound	(No Value)	Total
11-15	1.61%	1.40%	0.00%	3.00%
5-11	37.14%	40.26%	0.16%	77.56%
Less 5	6.78%	9.26%	3.15%	19.19%
Over 15	0.14%	0.12%	0.00%	0.25%
Total	45.66%	51.03%	3.31%	100.00%

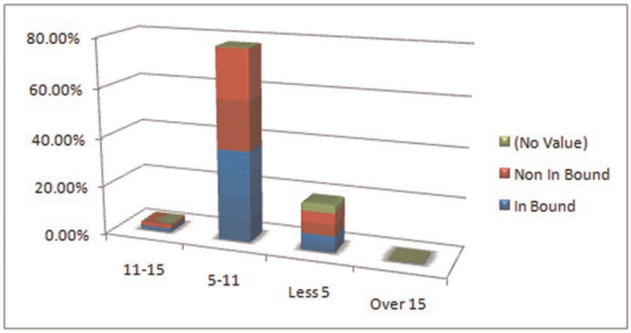

Figure 16. Draft and inbound/outbound pivot table.

InOut In Bound

%	Length				
Type	100-200	200-300	Less 100	Over 300	Total
Cargo Ship	61.42%	13.18%	7.08%	0.20%	81.89%
Passenger Ship	3.37%	0.56%	0.45%	0.00%	4.39%
Tanker	2.64%	0.00%	1.29%	0.00%	3.92%
Others	1.04%	0.23%	8.54%	0.00%	9.80%
Total	68.47%	13.98%	17.35%	0.20%	100.00%

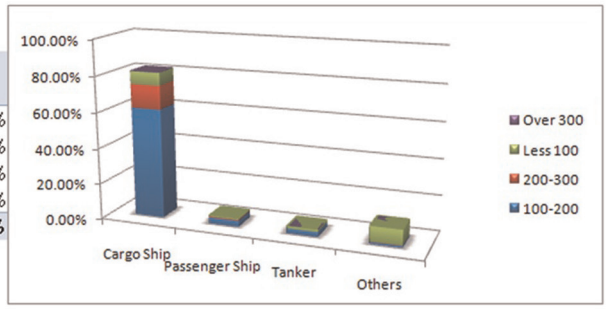

Figure 17. Ship length, ship type, and inbound/outbound pivot table.

%	In/Out		
Time	In Bound	Non In Bound	(No Value)	Total
⊞AM1	7.06%	13.82%	0.72%	21.60%
⊞AM2	11.70%	10.46%	0.65%	22.81%
⊟PM1	13.55%	11.06%	1.25%	25.86%
12	1.84%	1.46%	0.10%	3.41%
13	2.08%	1.67%	0.13%	3.88%
14	1.87%	1.81%	0.13%	3.81%
15	2.26%	1.94%	0.17%	4.37%
16	2.75%	2.05%	0.36%	5.16%
17	2.75%	2.13%	0.35%	5.22%
⊞PM2	13.36%	15.70%	0.68%	29.74%
Total	45.66%	51.03%	3.31%	100.00%

Figure 18. Temporal hierarchy.

integrates information systems that provide different types of text information but also other domestic or international VTS stations in related areas. Thus, it establishes an interconnected network of related marine traffic management systems. Due to this, it is foreseeable that the amount of data and complexity will increase, and a robust information tool will be required for data management. It is hoped that through this research's initial implementation, a foundation will be provided for establishing a future system that can conduct more complex processes over a wider area. It would also provide information to traffic management authorities to help them understand marine traffic

status and important information for marine traffic management decision-making.

Declaration of conflicting interests

The author declares that there is no conflict of interest.

Funding

The authors would like to thank the National Science Council, Taiwan, R.O.C., for financial support under contract number: 101-2410-H-022-009.

References

1. Harre I. AIS adding new quality to VTS systems. *J Navigation* 2000; 53(3): 527–539.
2. Naisbitt J. *Megatrends: ten new directions transforming our lives*. New York: Warner Books, 1982.
3. Mou JM, Tak C and Ligteringen H. Study on collision avoidance in busy waterways by using AIS data. *Ocean Eng* 2010; 37: 483–490.
4. Qu X, Meng Q and Suyi L. Ship collision risk assessment for the Singapore Strait. *Accident Anal Prev* 2011; 43: 2030–2036.
5. Silveira PAM, Teixeira AP and Guedes Soares C. Ship collision risks analysis off the coast of Portugal. *J Navigation* 2013; 66: 879–898.
6. Goerlandt F and Kujala P. Traffic simulation based ship collision probability modeling. *Reliab Eng Syst Safe* 2011; 96(1): 91–107.
7. Goerlandt F and Kujala P. On the reliability and validity of ship–ship collision risk analysis in light of different perspectives on risk. *Safety Sci* 2014; 62: 348–365, http://www.sciencedirect.com/science/article/pii/S0925753513002191
8. Montewka J, Hinz T, Kujala P, et al. Probability modelling of vessel collisions. *Reliab Eng Syst Safe* 2010; 95(5): 573–589.
9. Montewka J, Goerlandt F and Kujala P. Determination of collision criteria and causation factors appropriate to a model for estimating the probability of maritime accidents. *Ocean Eng* 2012; 40: 50–61.
10. Sormunen O-VE, Ehlers S and Kujala P. Collision consequence estimation model for chemical tankers. *Proc IMechE, Part M: J Engineering for the Maritime Environment* 2013; 227(2): 98–106.
11. Sormunen O-VE, Goerlandt F, Hakkinen J, et al. Uncertainty in maritime risk analysis: extended case study on chemical tanker collisions. *Proc IMechE, Part M: J Engineering for the Maritime Environment*. Epub ahead of print 2014. DOI: 10.1177/1475090213515640.
12. Ståhlberg K, Goerlandt F, Ehlers S, et al. Impact scenario models for probabilistic risk-based design for ship–ship collision. *Mar Struct* 2013; 33: 238–264.
13. Talavera A, Aguasca R, Galván B, et al. Application of Dempster–Shafer theory for the quantification and propagation of the uncertainty caused by the use of AIS data. *Reliab Eng Syst Safe* 2013; 111: 95–105.
14. Aarsæther KG and Moan T. Estimating navigation patterns from AIS. *J Navigation* 2009; 62: 587–607.
15. Zheng B, Chen J, Xia S, et al. Analysis of marine traffic flow characteristics based on data mining. *Navig China* 2009; 32(1): 60–63 (in Chinese).
16. Tsou M-C. Discovering knowledge from AIS database for application in VTS. *J Navigation* 2010; 63(3): 449–469.
17. Inmon WH. *Building the data warehouse*. Indianapolis, IN: John Wiley & Sons, 2005.
18. Kimball R, Ross M, Thornthwaite W, et al. *The data warehouse lifecycle toolkit: practical techniques for building data warehouse and intelligent business systems*. 2nd ed.Indianapolis, IN: John Wiley & Sons, 2008.
19. Goodwin EM. Marine encounter rates. *J Navigation* 1978; 31(3): 357–369.