Marine Traffic Engineering through Relational Data Mining

Antonio Bruno[1] and Annalisa Appice[1,2]

[1]Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro
via Orabona, 4 - 70126 Bari - Italy
[2]CILA (Centro Interdipartimentale per la ricerca in Logica e Applicazioni)
antonio.bruno@uniba.it,appice@di.uniba.it

Abstract. The automatic discovery of maritime traffic models can achieve useful information for the identification, tracking and monitoring of vessels. Frequent patterns represent a means to build human understandable representations of the maritime traffic models. This paper describes the application of a multi-relational method of frequent pattern discovery into the marine traffic investigation. Multi-relational data mining is here demanded for the variety of the data and the multiplicity of the vessel positions (latitude-longitude) continuously transmitted by the AIS (Automatic Identification System) installed on shipboard. This variety of information leads to a relational (or complex) representation of the vessels which by the way permits to naturally model the total temporal order over consecutive AIS transmissions of a vessel. The viability of relational frequent patterns as a model of the maritime traffic is assessed on navigation data truly collected in the gulf of Taranto.

1 Introduction

Marine traffic engineering is a research field originally defined in 1970s [10] at the aim of investigating the marine traffic data and building a human interpretable model of the maritime traffic. Through the understanding of this model, the Vessel Traffic Service (VTS) would improve the port and fairway facilities as well as the traffic regulation. Intuitively, the complexity of building a significant maritime traffic model resides in the requirement of a model able to reflect the spatial distribution and the temporal characteristics of the traffic flow.

Although, the marine traffic engineering was a popular research field between the 1970s and the 1980s, after the 1990s, the relevant literature and research projects in this field appeared less frequently. This little interest to research in marine traffic engineering was caused to the actual difficulty in collecting traffic data. In fact, the required observation time was long and several technological limitations raised in the observation time. Today, the data collection problem is definitely overcome. The widespread use of Automatic Identification System (AIS) has had a significant impact on the maritime technology and any VTS is now fit to obtain a large volume of traffic information which comprises the timestamped latitude and longitude of the monitored vessels. On the other hand,

the galloping developments in data mining research have paved the way to face the problem of automatically analyzing this large volume of traffic data, by the now available, in order to extract the knowledge required to feed the service marine traffic management and the VTS decision making systems. Both these factors, traffic data availability and data mining techniques, have boosted the recent renewed scientific interest towards the marine traffic engineering. Clustering [9], classification [5] and association rule discovery [11] techniques have been employed to analyze AIS data and discover characteristics and/or rules for the marine traffic flow forecast and the development and programming of marine traffic engineering. Although, these studies have proved that data mining techniques are able to provide the extra aid for the situational awareness in maritime traffic control, it is a fact that no marine traffic model described in these works is able to capture the truly temporal characteristics of each AIS transmission. In fact, AIS transmissions are timestamped, but a traditional data mining technique looses the time label of the AIS data, and represents a navigation trajectory as a set, rather than a sequence, of consecutive latitude-longitude vessel positions.

In this paper, we resort to multi-relational data mining to address the task of learning a human interpretable model of the maritime traffic in the sea ports, where several vessels are entering and leaving the port. The innovative contribution of this work is that, at the best of our knowledge, this is the first study in maritime traffic engineering which correctly spans the traffic data over several data tables (or relations) of a relational database and discover relational patterns (i.e. patterns which may involve several relations at once) to describe the traffic maritime model. In this multi-relational representation, we are able to model vessel data and AIS data as distinct relational data tables (one for each data type). This leads to distinguish between the reference objects of analysis (vessel data) and the task-relevant objects (AIS data), and to represent their interactions. The modeled interactions also include the total temporal order over the AIS transmissions for the same vessel. SPADA [6] is a multi-relational data mining method that discovers relational patterns and association rules. Relational patterns extracted by SPADA have been proved to be effective for the capture of the behavioral model underlying census data [1] and workflow data [12]. In the case of traffic data, we use SPADA to discover interesting associations between a vessel (reference objects) and a navigation trajectory. Each navigation trajectory represents a spatio-temporal pattern obtained by tracing subsequent AIS transmissions (task-relevant objects) of vessels. This kind of spatio-temporal rules automatically identify the well traveled navigation courses. This information can be employed in several ways. To opportunely arrange the navigation traffic incoming a gulf in order to avoid collision or traffic jams. To discover vessels which suspiciously deviates from the planned navigation course. The main limitation of SPADA in this application is the high computational complexity which makes the analysis of large databases practically unfeasible. To overcome this limitation, we run SPADA by considering the distributed version of SPADA described in [2].

In order to prove the viability of the multi-relational approach in marine traffic engineering, we describe a relational representation of the traffic data derived from monitoring vessels entered and left the gulf of Taranto (South of Italy) between September 1, 2010 (00:04:23) and October 9, 2010 (23:58:52) (Section 2) and we briefly illustrate the multi-relational method for relational pattern discovery (Section 3). Finally, we comment significance of navigation traffic model we have extracted and its viability in the marine traffic engineering (Section 4). Finally, some conclusions are drawn.

2 Marine Traffic Data

For this study, we consider the navigation traffic data collected for 106 vessels entering and/or leaving the gulf of Taranto between September 1, 2010 (00:04:23) and October 9, 2010 (23:58:52). The traffic data are obtained from [13]. As in [11], the area of the gulf is converted into a geographic grid of $0.005^\circ \times 0.005^\circ$ squared cells. Each cell of the grid is then enumerated by a progressive number. For each vessel, the following data are collected:

- the name of the vessel,
- the MMSI, that is, a numeric code that unambiguously identifies the vessel,
- the vessel category, that is, wing, pleasure craft, tug, low enforcement, cargo, tanker or other, and
- the sequence of AIS messages which were sent by the transceiver installed on shipboard.

The AIS transceiver sends dynamic messages every two to thirty seconds depending on the vessel speed, and every three minutes while the vessel is at the anchor. As we are interested in describing the observable change of the vessel position within the geographic grid, we decide to consider only those AIS transmissions which reflect a change of the cell occupied by the vessel. Each AIS message includes the following data:

- the vessel MMSI;
- the received time (day-month-year hour-minutes-seconds);
- the latitude and longitude of the vessel;
- the course over ground;
- the vessel speed;

The latitude and longitude coordinates of each AIS transmission are transformed into the identifier of the cell containing the coordinates. By following the suggestion reported in [11], the course over ground is discretized every 45° thus obtaining N, E, W, S, NE, NW, SE, SW, while the speed is discretized in low, medium and high. After this transformation, properties of vessels (name and category), data of the AIS transmission (cell, speed, direction) and interaction between vessel and transmitted AIS data are stored as ground atoms into the extensional part of a deductive database. An example of data stored in the

database for the vessel named ALIDA S is reported below.

mmssi(247205900).
name(247205900, ***alida s****).*
category(247205900, ***cargo****).*
ais(247205900, 2010-10-07 19:51:30).
ais(247205900, 2010-10-07 20:45:26).
ais(247205900, 2010-10-07 21:50:19).
ais(247205900, 2010-10-07 21:55:23).
cell(247205900, 2010-10-07 19:51:30, ***312****).*
cell(247205900, 2010-10-07 20:45:26, ***313****).*
cell(247205900, 2010-10-07 21:50:19, ***312****).*
cell(247205900, 2010-10-07 21:55:23, ***311****).*
direction(247205900, 2010-10-07 19:51:30, ***northwest****).*
direction(247205900, 2010-10-07 20:45:26, ***northwest****).*
direction(247205900, 2010-10-07 21:50:19, ***northwest****).*
direction(247205900, 2010-10-07 21:55:23, ***northwest****).*
speed(247205900, 2010-10-07 19:51:30, ***medium****).*
speed(247205900, 2010-10-07 20:45:26, ***medium****).*
speed(247205900, 2010-10-07 21:50:19, ***low****).*
speed(247205900, 2010-10-07 21:55:23, ***low****).*

The key predicate *mmsi*() identifies the reference object (vessel) of the unit of analysis. The property predicates *name*(), *category*(), *position*(), *direction*() and *speed*() define the value (in bold) taken by an attribute of an object (reference object as for *name*() and *category*() or task relevant object as for *position*(), *direction*() and *speed*()). Finally the structural predicate *ais*() relates reference objects (vessel) with task-relevant objects (AIS transmissions). This way, the extensional part of deductive database for SPADA is fed with 19137 atoms partitioned between 106 units of analysis.

3 Maritime Traffic Model Discovery

Studies for association rule discovery in Multi-Relational Data Mining [6] are rooted in the field of Inductive Logic Programming (ILP) [8]. In ILP both relational data and relational patterns are expressed in a first-order logic and the logical notions of generality order and of the downward/upward refinement operator on the space of patterns are used to define both the search space and the search strategy. In the specific case of SPADA, properties of both reference and task relevant objects are represented as the extensional part D_E of a deductive database D [4], while the domain knowledge is represented as a normal logic program which defines the intensional part D_I of the deductive database D.

In the application of SPADA in the marine traffic engineering, the extensional database stores information on the traffic data (e.g., vessel and AIS data) as reported in Section 2, while the intensional database includes the definition

of relations which are implicit in data, but useful for capturing the model underlying the data. In this study, the intensional part of database surely includes some definition of a relation *next* (which makes explicit the temporal order over the AIS transmissions that is implicit in the timestamp of each transmission). A possible definition of the relation *next* is the following:

next(V, A1, A2) ←
ais(V, T1), ais(V, T2),
cell(V, T1, A1), cell(V, T2, A2),
A1≠ A2, T1<T2,
not(ais(V,T), T1<T, T<T2)

which defines the direct sequence relation between two consecutive AIS transmissions of the same vessel.

In SPADA, the set of ground atoms in D_E is partitioned into a number of non-intersecting subsets $D[e]$ (unit of analysis) each of which includes facts concerning the AIS transmissions involved in a specific vessel trip e. The partitioning of D_E is coherent with the individual-centered representation of training data [3], which has both theoretical (PAC-learnability) and computational advantages (smaller hypothesis space and more efficient search). The discovery process is performed by resorting to the classical levelwise method described in [7], with the variant that the syntactic ordering between patterns is based on θ-subsumption. By SPADA, fragments of the traffic models underlying the navigations of the various traced vessels can be expressed in the form of relational navigation rules in the form:

$$mmsi(V) \Rightarrow \mu(V) \quad [s, c],$$

where $mmsi(V)$ is the atom that identifies a vessel, while $\mu(V)$ is a conjunction of atoms which provides a description of a fragment of the navigation trajectory traced for V. Each atom in $\mu(V)$ describes either the next relation between AIS transmissions or a property of the vessel (type or length) or a datum included in the AIS transmission (id of the crossed geographical cell, navigation direction, velocity). An example of discovered association rule is the following:

vessel(V)⇒
cell(V,T,123), next(V,123,124), next(V,124,125)
[s=63%, c=100%]

The support s estimates the probability $p(vessel(V) \bigcup \mu(V))$ on D. This means that s% of the units of analysis $D[e]$ are covered by $vessel(V) \bigcup \mu(V)$, that is a substitution $\theta = \{V \leftarrow e\} \cdot \theta_1$ exists such that $vessel(V) \bigcup \mu(V)\theta \subseteq D[e]$. The confidence c estimates the probability $p(\mu(V)|vessel(V))$.

Our proposal is to employ SPADA in order to process large traffic data volume and to collect the navigation rules discovered by SPADA in order to obtain an interpretable description of the model underlying the maritime traffic. As the

navigation rules describe fragments of the trajectories frequently crossed by the monitored vessels, they are then visualized in a GIS environment for the human interpretation. At the aim of this study, we have further extended SPADA by integrating a rule post-processing module which filters out uninteresting rules and ranks the output of the filtering phase on the basis of the rule significance. Then, the top-k rules compose the maritime traffic model. Interesting rules correspond to non-redundant rules. Formally, let R be the navigation rule set output by SPADA. A rule $r \in R$ is labeled as redundant in R iff there exists a rule $r' \in R$ and the substitution θ such that $r\theta \subset r'$. For example, let us consider the set of navigation rules which comprises:

r1: *vessel(V)⇒ cell(V,T,123).*
r2: *vessel(V)⇒ cell(V,T,123), next(V,123,124).*
r3: *vessel(V)⇒ cell(V,T,123), next(V,123,124), next(V,124,125).*

Both $r1$ and $r2$ are redundant in R due to the presence of $r3$.

Redundant rules are implicit in non-redundant rules (although, they may have different support, they are always frequent rules), hence we can filtered out the redundant navigation rules without loosing any knowledge in the maritime traffic model which is built finally. Filtered rules are ranked on the basis of significance expressed by pattern length (number of atoms in the rule and support value). By decreasing k, we prune less significant knowledge in the model.

4 Maritime Traffic Models

A relational model of the maritime traffic in the gulf of Taranto (South of Italy) was extracted by considering two experimental settings, denoted as S1 and S2. In the former setting (S1), the intensional part is populated with the definition of the ternary "next" predicate formulated as in Section 2. In the latter setting (S2), the intensional part is populated with an intensional definition of both a new "cell" predicate and a "next" predicate which incorporate the information on the speed and direction of navigation as follows:

cell(V, T, C, S, D) ←
cell(V, T, C), speed(V, T, S), direction(V, T, D).

next(V, A1, A2, S, D) ←
ais(V, T1), ais(V, T2),
cell(V, T1, A1), cell(V, T2, A2),
speed(V, T2, S), direction(V, T2, S)
A1≠ A2, T1<T2,
not(ais(V, T, A), T1<T, T<T2).

In both settings, SPADA is run to discover relational rules with 0.1 as minimum support and 3 as minimum pattern length. In the first setting, SPADA

outputs the geometrical description fragments of navigation trajectories incoming and leaving the gulf of Taranto. The number of discovered rules is 126. After filtering out redundant rules, 41 rules are ranked according to the significance criterion. The top ranked navigation rule is reported below:

vessel(V)⇒
category(V,cargo), cell(V, T, 903),
next(V, 903, 904), next(V, 904, 944),
next(V, 944, 945), next(V, 945, 946), next(V, 946, 986),
next(V, 986, 987). [s=10.3%, c=100%]

This rule states that 10.3% vessels monitored in the gulf of Taranto in the period under study are cargo vessels which follow a navigation trajectory crossing across the cells identified by 903, 904, 944, 945, 946, 986 and 987 in this order.

The maritime traffic model obtained by selecting the top-5 navigation rules is plotted in Figure 1. By visualizing this model we are able to see the geometrical representation of maritime trajectories which may be busy in the gulf of Taranto. This information may be employed from the service maritime management in order to opportunely program the maritime traffic in the gulf of Taranto and avoid gridlock or vessel accident.

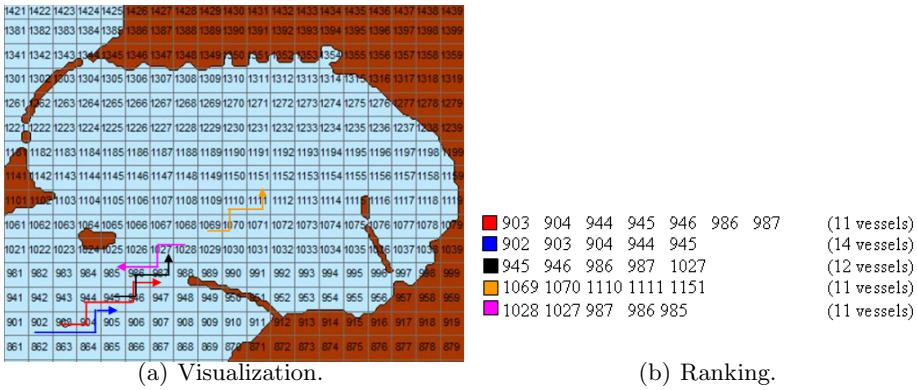
In the second setting, SPADA discovers a more detailed description of the navigation trajectories frequently crossed in the gulf. In fact, the description mined for each navigation trajectory now comprises both direction and velocity of the vessel at each crossed cell in the trajectory. With this setting, SPADA discovers 11 navigation rules. After filtering out redundant rules, 8 rules are ranked according to the significance criterion. The top ranked navigation rule is reported below:

vessel(V)⇒
category(V,cargo), cell(V, T, 945, low, north_east),
next(V, 945, 946, low, north_east), next(V, 946, 986, low, north_east).
[s=11.3%, c=100%]

This rule states that 11.3% of vessels in this study move across the cells 945, 946 and 986 maintaining a low velocity and north-east navigatition dierection. Although this navigation rule describes a shorter trajectory than the top ranked rule of the first setting, it provides a deeper insight in the navigation behaviour (velocity and direction) of vessels crossing these cells, which where ignored before.

5 Conclusions

In this paper, we presented a preliminary study of the application of relational data mining to the marine traffic engineering. Relational data mining is here demanded to represent multiplicity and variety of data continuously transmitting



(a) Visualization. (b) Ranking.

Fig. 1. The top-5 relational models of the incoming and outcoming navigation trajectories frequently crossed in the gulf of Taranto.

from a vessel during the navigation time. In particular, we prove the viability of a multi-relational approach to obtain human interpretable maritime models of the maritime traffic by considering the AIS data transmitted from vessels in the gulf of Taranto. The results are encouraging and open appealing and novel directions of research in the field of the marine traffic engineering.

As future work, we plan to explore the task of discovering relational rules which include a disjunction of atoms in the rule body in order to describe those trajectories which include one or more ramification in the path. Additionally, we intend to use the discovered navigation trajectories to obtain a prediction model that permits to predict the position of a vessel at any future time. This task requires the consideration of either geographical constraints such as the presence of the mainland (or in general physical obstacles) or navigation constraints such as velocity, direction, timetable and so on.

Acknowledgment

This work is partial fulfillment of the research objective of ATENEO-2010 project entitled “Modelli e Metodi Computazionali per la Scoperta di Conoscenza in Dati Spazio-Temporali”.

References

1. A. Appice, M. Ceci, A. Lanza, F. A. Lisi, and D. Malerba. Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis*, 7(6):541–566, 2003.
2. A. Appice, M. Ceci, A. Turi, and D. Malerba. A parallel, distributed algorithm for relational frequent pattern discovery from very large data sets. *Intell. Data Anal.*, 15(1):69–88, 2011.

3. H. Blockeel and M. Sebag. Scalability and efficiency in multi-relational data mining. *SIGKDD Explorations*, 5(1):17–30, 2003.
4. S. Ceri, G. Gottlob, and L. Tanca. *Logic Programming and Databases*. Springer-Verlag New York, Inc., New York, NY, USA, 1990.
5. R. Lagerweij. *Learning a Model of Ship Movements*. Thesis for Bachelor of Science - Artificial Intelligence, University of Amsterdam, 2009.
6. F. A. Lisi and D. Malerba. Inducing multi-level association rules from multiple relations. *Machine Learning*, 55(2):175–210, 2004.
7. H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.
8. S. Muggleton. *Inductive Logic Programming*. Academic Press, London, 1992.
9. C. Tang and Z. Shao. Modelling urban land use change using geographically weighted regression and the implications for sustainable environmental planning. In Q. Peng, K. C. P. Wang, Y. Qiu, Y. Pu, X. Luo, and B. Shuai, editors, *Proceedings of the 2nd International Conference on Transportation Engineering*, pages 4465–4470. ASCE, American Society of Civil Engineering, 2009.
10. S. Toyoda and Y. Fujii. Marine traffic engineering. *The Journal of Navigation*, 24:24–34, 1971.
11. M.-C. Tsou. Discovering knowledge from ais database for application in vts. *The Journal of Navigation*, 63:449–469, 2010.
12. A. Turi, A. Appice, M. Ceci, and D. Malerba. A grid-based multi-relational approach to process mining. In S. S. Bhowmick, J. Küng, and R. Wagner, editors, *Proceedings of the 19th International Conference on Database and Expert Systems Applications, DEXA 2008*, volume 5181 of *Lecture Notes in Computer Science*, pages 701–709. Springer, 2008.
13. web url:. http://www.marinetraffic.com/ais.