Efficient AIS Data Processing for Environmentally Safe Shipping

Marios Vodas[1], Nikos Pelekis[2], Yannis Theodoridis[1], Vangelis Karkaletsis[3], Sergios Petridis[3], Anastasia Miliou[4]

[1]Department of Informatics, University of Piraeus
Email: {mvodas, ytheod}@unipi.gr

[2]Department of Statistics and Insurance Science, University of Piraeus
Email: npelekis@unipi.gr

[3]Institute of Informatics and Telecommunications, NCSR "Demokritos"
Email: {vangelis, petridis}@iit.demokritos.gr

[4]Archipelago – Institute of Marine Conservation
Email: a.miliou@archipelago.gr

Abstract. Even though ship accidents at sea has many economic and environmental implications on Greece, ships formulate routes according to their best judgment. In this study we take a dataset spanning in 2.5 years from the AIS network, which is transmitting in public a ship's identity and location, and we load it in a trajectory database supported by the Hermes MOD system to begin an analysis by extracting safety related statistics for the dataset. Simple queries as well as more complex ones (e.g. OD-Matrix) on the dataset illustrates the capabilities of Hermes and allows to gain insight on how the ships move in the Greek Seas. One of the newest challenges that emerged during this process is that the amount of the positioning data is becoming more and more massive. As a conclusion, a preliminary review of possible solutions to this challenge along with others is mentioned.

1 Motivation

AIS data are massively available through internet sources (e.g. marinetraffic.com, mariweb.gr, aishub.net, vesseltracker.com) in the form of streams of messages, so called AIS sentences, and nearly at no cost. Nowadays, position reports in European waters reach almost 5 million per day corresponding to 17,000 ships (EMSA, 2012). Keeping in mind that the number of transceivers is continuously increasing and the quality of data is improving it is clear that the effective amount of data we have to process is ever increasing.

The amount of data gathered, together with their spatial/temporal distribution and their origin, provide further challenges related to distributed and scalable reasoning, retrieval of data from the different sources considered, “feeding” the knowledge-

intensive tasks of interest. Issues related to the provenance and trustworthiness of incoming data sources, which are of major importance to the safe shipping domain, should be also considered, mainly focusing on the less studied intertwined nature of the sources (Zhao et al. 2004).

The raw format for AIS data in relational database terms is a table of positions along with the timestamp each position was observed by the base stations. Querying that schema has certain drawbacks in terms of computing cost and interpolation. Consider a ship reporting positions around and inside an area of interest. It is very difficult and complex to determine if a ship entered or left this area using only the point-based model.

Trajectory features are required to allow us to inherently solve most issues that arise from movement of objects, such as interpolation and distance computation. Furthermore, by using this model we can reduce the computational cost by using simplification methods to limit the number of points in the trajectory.

The ways to define a trajectory vary significantly and are subject to the domain they are applied (Spaccapietra, 2008). Our focus in this study is mainly on answering questions regarding the movement of ships based on spatiotemporal predicates. Therefore, we model a trajectory as a sequence of sampled time-stamped locations (p_i, t_i) where p_i is a 2D point (x_i, y_i) and t_i is the recording timestamp of p_i. We can choose from two alternatives for interpolating the position of an object between two sampled points. The first and most common one is to assume constant speed linear interpolation and the second one is to consider constant acceleration. Although, the second option is closer to a real-world model we decided that the first option is more suitable for our data since a ship would not change its speed as rapidly as a car for example.

This need for capturing, storing and querying complete histories of objects' movement has promoted the investigation of Moving Object Databases (MOD) (Güting et al. 2000), which are based on the modelling of moving extensions of spatial types, such as points, lines, regions, etc. A state-of-the-art realization of the above model in Object-Relational DBMS is the so-called *Hermes* (Pelekis et al. 2008). Hermes provides an extensive spatiotemporal functionality over MODs, including aspects from online data feeding and efficient storage to trajectory similarity functions (L1, Euclidean, Linf, LCSS, DTW, ERP, EDR, dissim) and to advanced query processing and data mining.

2 Efficient AIS data processing

This section introduces our first real world experience with a massive AIS database. The database in its original form is almost 2 TB in size and contains records of raw AIS sentences from April 2007 to June 2010.[1] However, for the purposes of this study, we extracted a 3 days subset which contains approximately 3 million GPS records from 933 ships, spans in a 3 days period, from 31/12/2009 to 2/1/2010, and co-

[1] The "IMIS 3 Days" dataset is publicly available at ChoroChronos (http://www.chorochronos.org/?q=node/8).

vers an area of 496736 km^2 in a rectangle bounded by coordinates from (21, 35) to (29, 39) expressed in WGS84 (i.e. long, lat).

The following figure (Fig. 1) graphically presents the steps in our data processing methodology. These several steps are explained in detail throughout the next paragraphs.

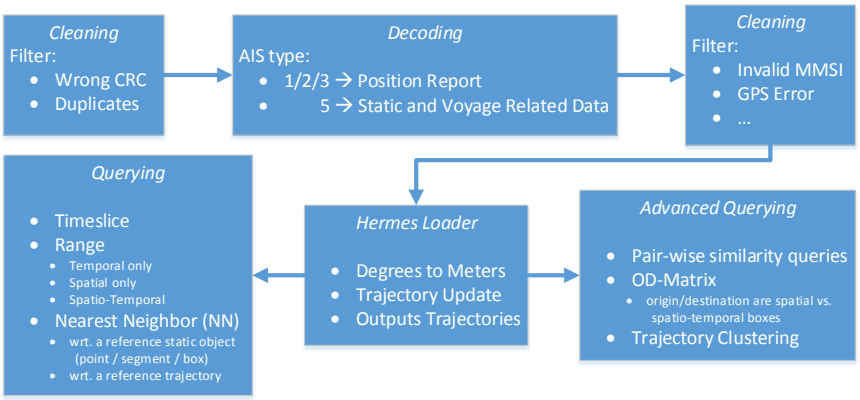

Fig. 1 Processing methodology

2.1 Loading the AIS data in a MOD

Examination of the data showed the first challenges in cleaning the data. Each AIS sentence is received multiple times from different stations. Another problem is that some sentences are incorrect, w.r.t. the checksum each sentence is accompanied with, or not whole therefore can't be decoded. It is expected to confront these problems in every raw AIS dataset (historical or streaming). In this case study, where the database is historical, cleaning duplicated and bad messages is straightforward and even in a centralized system the computational cost is acceptable. In order to speed this operation we partitioned data into multiple tables using the AIS checksums to make the cleaning process more effective. More ways can be used for partitioning of course but our experience in this experiment showed that choosing checksums as the partition key results in relatively same size partitions.

To extract the position reports from the AIS sentences we used Hermes's built-in AIS decoder and applied simple filters on the positions and the MMSI's of the ships to eradicate invalid ship identifiers and GPS errors. After that, the *Loader* module of Hermes is used to construct ship trajectories out of the decoded dataset. An index structure on the trajectories is also built to make queries with spatio-temporal predicates more efficient.

Understanding the dataset is prerequisite for its efficient and effective exploration. Typical statistical analysis, such as the number of ships and positions recorded, the spatial and temporal bounds of the dataset that were shown previously is essential, although, it is not as powerful as the kind of analysis we present in the next paragraphs.

2.2 Querying the AIS database, focusing on spatio-temporal attributes

Taking geographic information into account is of paramount interest in processing an AIS database. Apart from the information already embedding in AIS signals, complementary information can be obtained from official maritime vector charts that (eg. S-57) contain different kind of objects useful for spatial analysis (coast lines, restricted areas …) or even added manually, according analysis objectives. In our study, we have integrated geographic data, such as points or regions of interest (ports, lighthouses), coastal boundaries, etc. Thereafter, we were able to support queries, such as:

- Find the ships that crossed a narrow passage;
- Find the ships that approached (closer than a certain distance) a point of interest;
- Calculate the number of sharp changes in ship's direction;
- etc.

Hermes provides the ability to build a 3D-Rtree index on the spatio-temporal segments of the trajectories in order to efficiently process queries in this category (Theodoridis, 1996). This tree structure is implemented on top of GiST interface (Hellerstein, 1995) and can support the retrieval of spatiotemporal objects based on temporal, spatial, and spatiotemporal predicates. Basically, the 3D-Rtree has an operator class associated with it so that every time we use an operator from that class we can take advantage of the index. An example of such operators would be the "within distance" which will traverse the tree and will return only the segments that fulfill the distance threshold we will give. This operator was used in the query shown in Fig. 2. Other examples include the "intersects" and "contains" operators.

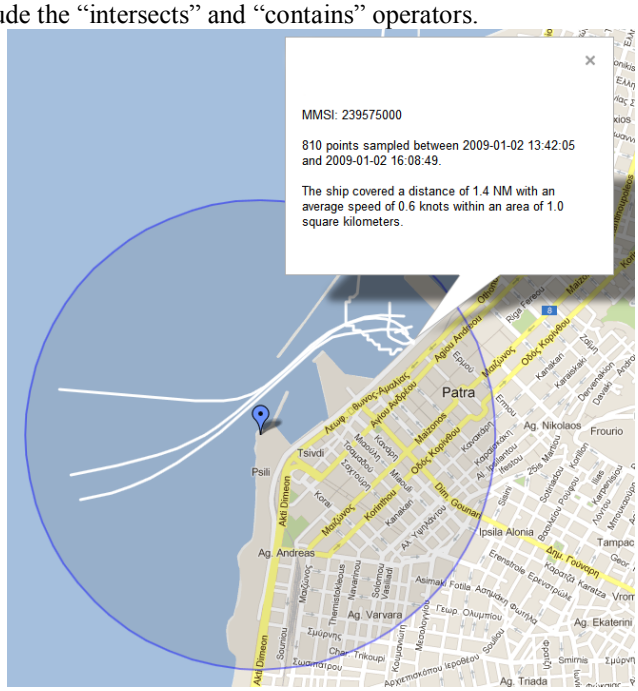

Fig. 2 Ships located closer than ½ n.m. from an old lighthouse in port of Patras

2.3 Advanced processing of vessel routes

The queries belonging this category are more complex and cost more than the previous ones. Yet, they are very interesting and provide deeper insight into the data. Some of the complex operations that we are able to support are:

- Building OD-matrix between areas of particular interest (Table 1);
- Clustering trajectories using T-Optics or S^2T-Clustering algorithms;
- Find typical routes and compare ship trajectories to them;
- etc.

As shown by the OD-Matrix (Table 1) the two main entrance and exit routes in Greek Seas are the most heavily used routes. Meaning that Greek Seas are a congested crossroad for travelling ships. A further simple statistical analysis on the dataset shows that cargo ships are the majority and that there is a large number of flags of convenience, therefore we can conclude that national monitoring and the formation of safe routes are necessary in order to prevent an accident that could cause irreversible damage both to the ecosystem and the economic driving force of Greece, tourism.

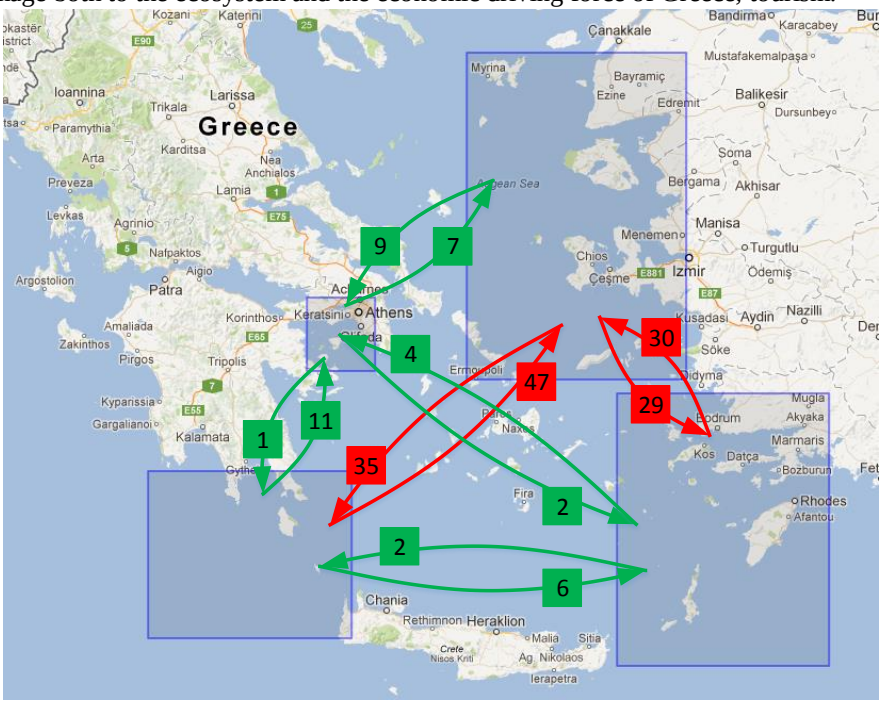

Fig. 3 Origin-Destination Matrix between 4 large entrance and exit areas of the Greek territory

3 Conclusion

Maritime environment represents an increasing potential in terms of modelling, management and understanding of mobility data. Recently, several real-time positioning

systems, such as the Automatic Identification System, have been developed for keeping track of vessel movements. Analysis of such positioning data needs cooperation and input from experts in environmental and marine domains. Indeed, knowing the everyday actions of ships and their patterns help to model them and put them to efficient use. Case-based analysis with experts also provides a better understanding of the problems and challenges and guides researchers towards efficient data mining process.

This research aims at exploring new ways to analyse and visualize the information gathered and publish it to stakeholders and the public. Automatically processing huge amounts of such data in real time is a very challenging task which, when addressed, will give the ability to the public and stakeholders to report environmental wrongdoing of ships.

Acknowledgements

This research was partially supported by AMINESS project funded by the Greek government. Cyril Ray was supported by a Short Term Scientific Mission performed at the University of Piraeus by the COST Action IC0903 on "Knowledge Discovery from Moving Objects" (COST-MOVE, http://www.move-cost.info). IMIS Hellas (www.imishellas.gr) kindly provided the AIS dataset for research purposes.

References

Devogele T., Etienne, L., Ray, C., Mobility Data: Modelling, Management, and Understanding, Part 3, Chapter 11 : Maritime monitoring, pages 224-243, Chiara Renso, Stefano Spaccapietra, Esteban Zimanyi (eds), 2013, Cambridge University Press.

EMSA, 2012, European Maritime Safety Agency, Annual Report 2011, 114 pages, June 2012.

Güting RH, Böhlen MH, Erwig M, Jensen CS, Lorentzos NA, Schneider M and Vazirgiannis M, 2000, A foundation for representing and querying moving objects. *ACM Transactions on Database Systems*, 25(1):1–42.

Hellerstein J, Naughton J, and Pfeffer A, 1995, Generalized Search Trees for Database Systems. In *Proceedings of VLDB*, 562-573.

Pelekis N, Frentzos E, Giatrakos N and Theodoridis Y, 2008, HERMES: aggregative LBS via a trajectory DB engine. In *Proceedings of ACM SIGMOD*.

Spaccapietra S, Parent C, Damiani M, Macedo J, Porto F, and Vangenot C, 2008, A conceptual view on trajectories. *Data & Knowledge Engineering*, 65(1):126-146.

Theodoridis Y, Vazirgiannis M, and Sellis T, 1996, Spatio-Temporal Indexing for Large Multimedia Applications. In *Proceedings of ICMCS*, page 0441.

Zhao J, Wroe C, Goble C, Stevens R, Quan D and Greenwood M, 2004, Using semantic Web technologies for representing E-science provenance. In *Proceedings of ISWC*, 92-106.