Applying Knowledge Discovery and Semantics for Detecting Anomalous Container Trips

Elena Camossi[1], Tatyana Dimitrova[2], Luca Mazzola[2], Aris Tsois[2], Paola Villa[2]

[1]elena.camossi@gmail.com
[2]{tatyana.dimitrova, luca.mazzola, aris.tsois, paola.villa}@jrc.ec.europa.eu
European Commission Joint Research Centre (JRC)
Institute for the Protection and Security of the Citizen (IPSC)
Ispra, Varese, Italy

Abstract. Containers have become one of the most concrete breach for global security, and there is a pressing need of novel methods to target highly risky container shipments. In this paper we present the ongoing research carried out at the European Commission Joint Research Centre for the development of data mining and semantic approaches for route based risk indicators to discover anomalous container trips.

1 Introduction

About 20 millions of containers per year travel by sea, transporting around the 25% of the world's trade. Physically inspecting their content is a long and expensive operation, but other types of security checks (e.g., X-rays, scanning, content sniffing) are not as effective; moreover, from many parts it is argued that increasing inspections would affect negatively the world trade market. This situation paves the way to leverage on containers to pursue a number of illicit activities, from frauds to avoid Communitarian antidumping duties and circumventing import/export quotas, to counterfeited goods smuggling, up to security violations such as smuggling prohibited chemicals, nuclear materials and weapons, up to use them to accomplish a terroristic attack. Indeed, after 9/11 containers have been identified as one of the most relevant breach for global security, making supply chain security a major issue. Therefore, it is necessary to improve the capacity of Customs authorities of targeting high risk consignments and timely identifying and disarming any risky situation.

ConTraffic is a global route-based Risk Analysis facility for maritime containers monitoring developed at JRC. The ConTraffic system collects and analyses historical information on containers and cargo vessels travelling worldwide, and develops risk indicators to identify irregular and anomalous shipments and transportations. The system findings are signalled to Customs Authorities, helping overcome ineffective random inspections and improve the efficacy of security checks.

The base data collected by the system are Container Status Messages (CSM), text messages shipping companies use for exchanging logistic information that

describe the operations carried out on containers and their position, namely the *container events* (e.g., loaded, discharged, transhipped). Each CSM describes a deed occurring to one container (univocally identified by an ISO 6346:1995 code [4]), and includes also a timestamp, the container loading status, and sometimes, the identifier of the vessel used for transportation. CSM are semi-structured and not standardised, therefore an intensive cleaning process is required to remove text typos and errors, to reduce inconsistencies and uncertainty before being analysed [7]. Afterwards, for each container in the database, the CSM sequence, namely, the *container history*, is considered. A container history can be segmented in several *trips*, each representing the *route*, or the trajectory followed by a container to fulfil a shipment, i.e., from the initial stuffing operation to the final consignment of the goods.

On CSM and trips risk indicators are defined and analysed according to stakeholders requests and needs. In this paper, we describe two research projects recently undertaken in ConTraffic to discover *anomalous* container trips. The first project, described in Section 2, exploits Knowledge Discovery, specifically unsupervised anomaly detection [1], to identify shipments that are outliers with respect to the dataset's distribution. By contrast, the second project, which is presented in Section 3. applies Semantic Technologies, in particular semantic reasoning and queries, to search for anomalous container trip patterns formalised within an ontology. The paper concludes outlining future research directions.

2 Knowledge Discovery of Anomalous Container Trips

The Anomalous Container Itineraries (trips) Detection (ACID) prototype discover suspicious container shipments by applying one-class classification, an *unsupervised* approach that marks the outliers in a distribution estimated directly on the test dataset. Unsupervised approaches are appropriate in Risk Analysis, where counter-examples needed by supervised approaches are rarely available and usually unprofitable because new risk patterns continue to emerge. Specifically, ACID uses Support Vector Machines (SVM) [2], a well known kernel-based linear classification algorithm, and focuses the classification on aggregated spatio-temporal features and on geographical information extracted from trips.

The analysis module of the prototype extends the open source Weka Data Mining toolkit API [13] and the SVM API LibSVM [6]. In the current version of the prototype, the core API for SVM has been extended to geographical kernel functions and to multi-kernel functions, i.e., combined kernel functions that handle input datasets with heterogeneous data types (specifically, vectors of numbers, strings and geographical coordinates).

Given a CSM dataset, ACID selects the corresponding container histories and pre-processes the data defining the analysis *features* of interest, which correspond to risk indicators. In particular, the ACID data preparation algorithm iteratively segments CSM sequences into container trips, relying on the semantics of container events specified within the CSM and on the changes of the loading status of the container to identify the different phases of a shipment and

to distinguish one container trip from the following one. The features extracted at this step include typical spatio-temporal risk indicators, such as the time to accomplish the shipment, the time to complete the maritime trip, the number of transhipments from a vessel to another, the origin and the destination of the shipment, the ports involved during the trip.

ACID runs the classification algorithm on the extracted features and finds a preliminary set of anomalies; afterwards, a filtering step is applied, selecting a final set of suspicious trips that will be brought to the attention of Customs authorities for post-clearance processing. Result validation run by Customs is facilitated by the web-based geographical application provided with the system, namely, ACIDVIS, which maps anomalous trips against the most frequent routes followed to accomplish the same shipments (i.e., with the same origin and destination). In Fig. 1, a screenshot of the application is shown: ACID discovered an anomalous trip starting from Abidjan (CI) and ending to Alger (DZ), that includes two transhipments in Las Palmas (ES) and Valencia (ES). The anomalous trip is visualised in red in ACIDVIS, while the normal route between these two ports is depicted in blu.

The ACID prototype has been evaluated against a real world dataset containing more than 18 millions CSMs, which refer to 50 thousand containers travelling worldwide along a period of 4 years (from January 2009 to January 2013), testing RBF kernel and the newly developed geographical kernel included in the prototype.

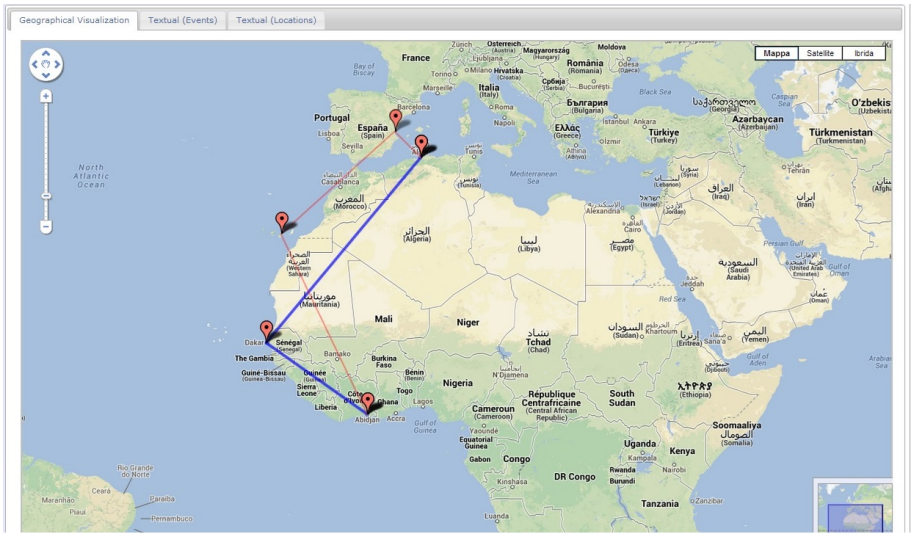

Fig. 1. ACIDVIS: example of anomalous container trip (in red). In blue, the most frequent route

3 Semantic Search of Anomalous Container Trips

In the effort towards developing novel methodologies for detection anomalous container trips, we have investigated the potentialities of ontologies and Description Logics (DL). The main result of this project is the methodology for the definition of Semantic Route-based risk Indicators (SemRI), ontology axioms representing *anomalous container trip patterns*. SemRI can be evaluated against a dataset of container trips represented in an ontology, exploiting both asserted and inferred knowledge on container events and vessel events, detecting anomalous trips.

In our test bed application, we have defined the Maritime Container Ontology (MCO), formalizing container and vessel events and trips, using OWL 2.0, the Web Ontology Language. Then, using ontology DL axioms, we have formalised two container movement patterns already tested in ConTraffic, which describe anomalous behaviour, namely, *cycle* and *unnecessary transshipment* [12]. We have populated the MCO with the dataset of container trips used to test ACID, enriched with information on vessel trips inferred by aggregating CSM data. To improve the performance of SemRI evaluation, these have been also implemented as semantic queries, formalised using OWL API [5] and the SPARQL-DL [10] syntax, and evaluated against three different reasoners: HermiT [8], Pellet [9] and Fact++ [11]. The combined use of SPARQL-DL and Pellet, together with customised temporal restriction on the evaluation of the queries, outperforms other reasoners and languages.

4 Conclusions

In this paper we have presented two experimental approaches that can be used to detect suspicious container shipments, adopting anomaly detection implemented with classifiers and semantic technologies. In both cases, an important issue to consider is the scalability of the proposed solutions. In ConTraffic, 30 millions of CSM are collected every months, therefore the size of the dataset can easily overcome the possibilities of traditional KD approaches and the limitation of OWL 2.0. A partial solution to this issue is the application of feature selection strategies. However, the number of features that can be analysed by a Risk Analysis system, can be enormous. Therefore, investigating incremental on-line analysis methods for high-dimensional data, as well as BIG DATA approaches, is a fundamental development for the system.

References

1. Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Survey*, 41(3):1–58, 2009.
2. Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

3. Tatyana Dimitrova, Aris Tsois, and Elena Camossi. Visualization of container movements through a web-based geographical information system. *(submitted)*, 2013.
4. International Organization for Standardization. Freight Containers - Coding, Identification and Marking, 1995.
5. M. Horridge and S. Bechhofer. The OWL API: A Java API for OWL Ontologies. *Semantic Web Journal, Special Issue on Semantic Web Tools and Systems*, 2:11–21, 2011. (accessed in December 2012).
6. LIBSVM - A Library for Support Vector Machines.
7. Luca Mazzola, Tatyana Dimitrova, Elena Camossi, Aris Tsois, Alberto v. Donati, and Mauro Pedone. Resolution of geografical string names. In *Proceeedings of the COST Move Workshop on Moving Objects at Sea, Brest, France*, June 2013.
8. R. Shearer and B. Motik and I. Horrocks. HermiT: A Highly-Efficient OWL Reasoner. In Alan Ruttenberg, Ulrile Sattler, and Cathy Dolbear, editors, *Proceedings of the 5th International Workshop on OWL: Experiences and Directions (OWLED 2008 EU)*, Karlsruhe, Germany, October 2008.
9. E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur, and Y. Katz. Pellet: A practical OWL-DL reasoner. *J. of Web Semantics*, 5(2):51–53, June 2007.
10. Evren Sirin and Bijan Parsia. Sparql-dl: Sparql query for owl-dl. In *Proceedings of the 3rd International Workshop on OWL: Experiences and Directions (OWLED-2007)*, 2007.
11. D. Tsarkov and I. Horrocks. FaCT++ Description Logic Reasoner: System Description. In *Proceedings of the International Joint Conference on Automated Reasoning (IJCAR 2006)*, volume 4130 of *Lecture Notes in Artificial Intelligence*, pages 292–297. Springer-Verlag, 2006.
12. P. Villa and E. Camossi. A Description Logic Approach to Discover Suspicious Itineraries from Maritime Container Trajectories. In *Proceedings of the 4th International Conference on GeoSpatial Semantics, (GeoS 2011), Brest, France, May 12-13, 2011*, pages 182–199, 2011.
13. Weka 3 - Data Mining with Open Source Machine Learning Software.