A Quality Assessment Framework for Large Datasets of Container-Trips Information

Michail Makridis, Raúl Fidalgo-Merino(✉), José-Antonio Cotelo-Lema, Aris Tsois, and Enrico Checchi

European Commission, Joint Research Centre (JRC), Ispra, Italy
{michail.makridis,raul.fidalgo-merino,antonio.cotelo, aris.tsois}@jrc.ec.europa.eu, enrico.checchi@ec.europa.eu

Abstract. Customs worldwide are facing the challenge of supervising huge volumes of containerized trade arriving to their country with resources allowing them to inspect only a minimal fraction of it. Risk assessment procedures can support them on the selection of the containers to inspect. The Container-Trip information (CTI) is an important element for that evaluation, but is usually not available with the needed quality. Therefore, the quality of the computed CTI records from any data sources that may use (e.g. Container Status Messages), needs to be assessed. This paper presents a quality assessment framework that combines quantitative and qualitative domain specific metrics to evaluate the quality of large datasets of CTI records and to provide a more complete feedback on which aspects need to be revised to improve the quality of the output data. The experimental results show the robustness of the framework in highlighting the weak points on the datasets and in identifying efficiently cases of potentially wrong CTI records.

Keywords: Quality assessment · Knowledge validation · Qualitative indicators · Supply chain

1 Introduction

The vast majority of non-bulk cargo worldwide is transported in containers. The World Shipping Council, estimates that more than 18 million containers were active in 2011 [5], transporting in 2014 more than 171 million TEU (Twenty-foot Equivalent Unit) of goods [12]. Due to the high volumes of containerized trade, with big ports handling more than 80000 TEU a day [12], authorities can physically check only a small fraction of it, limiting their capacity to detect illegal activities and security threats. To mitigate the risks, authorities focus on obtaining high quality information on the transported cargo to conduct effective risk assessments.

The 2015 WCO Safe Framework of Standards [13], defines as high-risk the cargo for which either there is inadequate information or reason to deem it as low risk, or is indicated as such by tactical intelligence or a risk-scoring assessment

K. Saeed and W. Homenda (Eds.): CISIM 2016, LNCS 9842, pp. 729–740, 2016.
DOI: 10.1007/978-3-319-45378-1_63

methodology based on security-related data elements. Customs administrations should ensure the interoperability of their IT systems by using the WCO Data Model. This data model identifies the key information elements for the Customs risk-scoring assessment methodologies and systems. Among them, WCO includes the container id, country of origin, place of loading, countries of routing, first port of arrival, place of discharge and date/time of arrival to the first port in Customs territory. The reason is that for goods shipped in containers, their route is equivalent to the container route since they are stuffed into the container (origin of the goods), until they are stripped out of the container (final destination of the goods). Unfortunately, poor quality information regarding the actual route followed by a container is quite common in Customs' declarations.

The logistics' industry and ocean carriers have been using for quite some time electronic records on the events and the whereabouts of the containers they handle. These records, called Container Status Messages (CSM) in WCO SAFE [13], are generated by different participants on the logistics chain (mainly container terminals) and are exchanged electronically to inform carriers, operators and final customers. Unfortunately, this information is often noisy, incomplete and non-standardized, requiring elaborated algorithms to remove the noise and improve the quality of the information. The works in [2–4,18] focus on how CSM records can be processed in order to extract useful information on the routes of the goods and assist the Customs risk management processes. The key element proposed on them to describe the route of the goods is the Container-Trip Information (CTI) [4,18]. In [4] an algorithm is presented for the computation of CTI records from CSMs based on Conditional Random Fields. Other algorithms, including decision trees, can also be applied to compute the CTI records.

The value of CTI records for Customs risk management is indisputable but it heavily depends on their quality. They are used to detect anomalies in the flow of millions of containers, so the quality of the CTI records extracted can impact severely the effectiveness of the risk analysis. Therefore, it is very important to be able to evaluate the quality of the models used to extract those CTI records. The final aim is to reduce the number of wrongly computed CTI records and their impact on the quality of the risk assessment, allowing a more efficient usage of the resources dedicated to container inspections by the authorities.

This paper proposes a CTI Quality Assessment Framework (CTI-QAF) that contributes to the improvement of the CTI record computation models and facilitates their evaluation. Based on quantitative and qualitative metrics, it provides both a way to evaluate the quality of the CTI computation model output and a way to identify the potentially wrong CTI records. Moreover, the output of the CTI-QAF facilitates the user to identify the CTI properties which are not correctly computed by the model.

The paper is organized as follows. Section 2 describes the related work. Section 3 formalizes the CTI representation to be used and describes the proposed CTI-QAF framework. Section 4 shows how this framework has been applied to assess the quality of four different case studies. Finally, Sect. 5 concludes the paper and highlights the main future lines.

2 Related Work

In this work we propose a quality assessment framework for automatically calculated CTI records (CTI-QAF). CTI-QAF provides useful information to improve the assessed CTI computation model and aims at facilitating risk assessment procedures by highlighting potentially wrong computed CTI records.

To be able to apply proposals like [6,19] to use CTI in Route-based Risk Indicators (RRIs), there is the need to develop algorithms to obtain this information from commonly existing data sources. In [2], CSMs were used to infer CTI records following a decision-tree like process. In [17] basic information on the container trip (origin, first port, last port and destination) is extracted from bill of lading documents [9]. In [4] a more sophisticated approach is proposed, extracting information on the different stages of a container trip from CSM data using Conditional Random Fields (see also [8] or [16]).

The assessment of the results in the above-mentioned techniques is based on traditional metrics of performance that focus on the generic accuracy of the proposed algorithm. They are not focused on the particularities of the domain and they cannot detect individual problematic cases. In other domains, like in the Part-of-the-speech tagging problem [11], one can find the usage of quality indicators. In this case, it is useful to know not only the precision of the algorithms when tagging words but also the *decision* (understood as the number of words non-ambiguously tagged [1]) or simply the sentence accuracy. Other examples of quality indicators can be also found in the Information Retrieval domain, as it is important to measure the document rank in web searches [7,10]. Such approaches are efficient in evaluating the overall quality of the information retrieval/extraction process, but they are not designed to evaluate the quality of the resulting data for outlier-detection environments.

In contrast, this paper proposes a new operational framework (CTI-QAF) that uses the CTI key element to formalize the goods route and proposes a set of quantitative (generic) and qualitative (domain specific) metrics to assess the overall quality of the model, the CTI properties and the individual CTI records.

3 CTI-Quality Assessment Framework

This section describes the proposed CTI-QAF, which is not only able to assess the quality of CTI datasets based on domain-specific metrics, but also capable to highlight potentially problematic CTI records. The outputs of the proposed CTI-QAF help the user to evaluate qualitatively the CTI computation model and to fine tune it by minimizing the number of wrongly computed CTI records.

3.1 CTI Formalization

A CTI record codifies the key route information regarding the transportation of goods in a container from an initial location where the goods were stuffed (inserted) into the container till the final destination where the goods were

stripped (extracted) from the container. The CTI record splits the route in 5 phases:

- Stuffing phase, when and where got the goods stuffed in the container?
- First-load phase, when and where started its maritime transport?
- Transshipment phase, when and where was it transshipped (if any)?
- Final-discharge phase, when and where ended its maritime transport?
- Stripping phase, when and where got the goods stripped from it?

For each phase CTI encodes: (a) the main location(s) involved, (b) the time period covered by the phase and (c) the vessel(s) involved (if any).

Each CTI can be described using the following representation:

$$cti = (containerId, stuffing, loading, transship[], discharging, stripping)$$

identifying the container and collecting the relevant data of each phase:

- *containerID* is the id that uniquely identifies a container box
- *stuffing*=(startDate, endDate, location)
- *loading*=(startDate, endDate, location, vessel)
- *transship[]* is a (possibly empty) list of records in the form: $transship_i = (startDate, endDate, location, vesselIn, vesselOut)$, with $i \epsilon [1, n], n \geq 1)$
- *discharging*=(startDate, endDate, location, vessel)
- *stripping*=(startDate, endDate, location)

We consider that a CTI record is wrongly calculated if a domain expert based on the same information would not conclude the exact same CTI record, even if the difference is in a single field of the CTI record.

Finding the potentially wrong CTI records out of a huge CTI set is not a trivial task when the computation model has a high precision. Random sampling and manual evaluation by domain experts (ground-truth) is not efficient as one would need to evaluate a very large number of CTI records before identifying enough wrong cases. The framework proposed addresses this problem by using qualitative metrics to identify the set of potentially wrong CTI records and using quantitative metrics to validate the quality of the selected dataset.

3.2 The CTI-QAF

The quality assessment upon a large result-set of any model is usually performed by defining one or more ground-truth subsets, which are then compared with the corresponding elements from the large results-set using quantitative metrics. The overall performance of the model can be then approximated by extrapolation. This type of quality assessment can be adequate in cases where an overview of the model's precision is enough for evaluation purposes. However, in cases such as the one discussed in this paper, we must be able to evaluate the expected impact that those wrongly computed CTI records will have on the risk assessment process.

The proposed CTI-QAF framework provides an overall quality assessment based on domain specific qualitative metrics, and additionally provides insight on which aspects need fine-tuning on CTI record computation models.

The workflow, depicted in Fig. 1, can be described as follows:

Input: A set of CTI records, *CTIset*, based on a specified computation model.

Stage A. Do a quantitative analysis on a small sample of *CTIset*.
- A.1 Select randomly a set of CTI records (we call that set *CTI-Xset*).
- A.2 Let experts create the ground-truth for *CTI-Xset* (*CTI-GT-Xset*).
- A.3 Perform quantitative analysis by computing *FDR* on the complete CTI records and *Precision*, *Recall* and *F1-Score* on CTI records' phases using *CTI-Xset* and *CTI-GT-Xset* (*QT-Xresults*).
- A.4 In case the *FDR* passes a predefined *FDRth* threshold, further improvement of the model is recommend.

Stage B. Perform the qualitative analysis on the entire CTIset.
- B.1 Apply the qualitative metrics *QLM* to the *CTIset* and extract a set of potentially wrong CTI records, *pwrongCTIset*.
- B.2 Order decreasingly the *pwrongCTIset* based on how many metrics signaled each CTI record, producing *RListpwrongCTIset*.
- B.3 For each metric, calculate the percentage (QLM_i) of the CTI records evaluated with it that have been detected as suspicious. If for any indicator the percentage is higher than a predefined *QLTh* threshold, the further improvement of the model is recommend.

Stage C. Do a quantitative analysis on a small sample of *RListpwrongCTIset*.
- C.1 Select the set of CTI records with highest risk according to the *RListpwrongCTIset* and call that set *CTI-RXset*.
- C.2 Let experts create the ground-truth for *CTI-RXset* (*CTI-GT-RXset*).
- C.3 Perform quantitative analysis by computing *FDR* on the complete CTI records and *Precision*, *Recall* and *F1-Score* on CTI records' phases using *CTI-RXset* and *CTI-GT-RXset* (*QT-RXresults*). The results can show: (1) if the qualitative metrics properly select the *RListpwrongCTIset* set and, (2) which are the weakness of the CTI computation model if any.

The next section describes the metrics used in the proposed CTI-QAF.

3.3 Quantitative and Qualitative Metrics

This section describes in detail the quantitative and qualitative metrics used in the framework in order to assess the input CTI dataset.

Quantitative Metrics. The metrics most commonly used to measure the performance of classification models are based on confusion tables by comparison of the classifiers' results with the ground-truth on a sample set [15]. In CTI-QAF we use *Precision*, *FDR*, *Recall* and *F1-Score*:

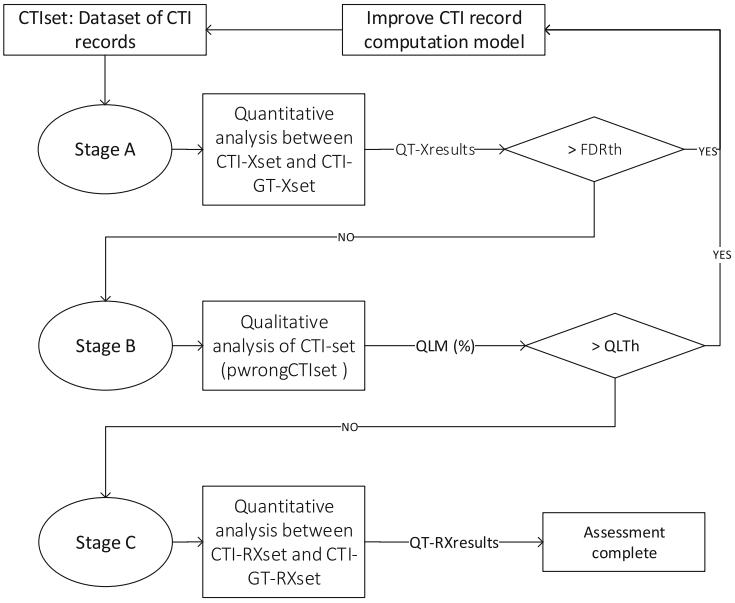

Fig. 1. Workflow of the CTI-QAF

- *Precision* is the success rate on those elements for which the model assigned a positive class and it is defined as: $Precision = \frac{TP}{TP+FP}$
- *False discovery rate* (FDR) is the opposite of precision: $FDR = 1 - Precision$
- *Recall* is the coverage of real positive elements achieved by the model and it is defined as: $Recall = \frac{TP}{TP+FN}$
- *F1-Score* provides combined information about the *precision* and *recall* of the results obtained from the model and it is defined as: $F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$

where: TP is the number of True Positive (the classifiers resulted class coincides with the ground-truth), FP is the amount of False Positives (the classifier does not coincide with the ground-truth) and FN stands for False Negatives (classifier's results that are not included in the ground-truth result set).

Qualitative Metrics. We propose to use several domain-specific metrics aimed to assess the quality of the CTI record computation model.

We split the proposed metrics into heuristic metrics and semantic indicators. The first are based on a set of threshold functions that aim to detect CTI records outliers based on deviations from typical container trip characteristics. The second target semantically incomplete or potentially wrong CTI records, i.e., container trips or trip-phases having semantically incoherent properties.

The heuristic metrics defined below are based on the duration of the different parts of the trip and the number of transshipments.

– *Abnormally long trips*: Trips with duration equal or greater than 120 days.

$$TLong = (cti.stripping.endDate - cti.stuffing.startDate \geq 120)$$

– *Abnormally short trips*: Trips with duration equal or shorter than 5 days.

$$TShort = (cti.stripping.endDate - cti.stuffing.startDate \leq 5)$$

– *Trips with multiple transshipments*: Trips more than 3 transshipments.

$$TMulti = (|cti.transship| > 3)$$

– *Trips with prolonged initial phase*: Trips having cumulative duration of stuffing and first-load phases equal or greater than 20 days.

$$TSDLong = (cti.loading.endDate - cti.stuffing.startDate \geq 20)$$

– *Trips with prolonged final phase*: Trips having cumulative duration of final-discharge and stripping phase sequal or greater than 20 days.

$$TARLong = (cti.stripping.endDate - cti.discharging.startDate \geq 20)$$

The semantic indicators defined below assess the structure and quality of the CTI records through functions that highlight semantic inconsistencies.

– *Trips with same location for load and discharge*: Container trips having the same location as origin and destination.

$$TLocation = (cti.stuffing.location == cti.discharging.location)$$

– *Trips with at least one transshipment and having the same vessel at the loading and the discharging phases*: During the container trip several transshipments may take place, i.e. the containers are being discharged from the vessel and re-loaded to a different vessel.

$$TODVessel = ((cti.loading.vessel == cti.discharging.vessel) \wedge (|cti.transship| \geq 1))$$

– *Trips joined with SAD declarations having different origin*: In the EU, the Single Administrative Document (SAD) [14] is used in the trade with third countries and for moving non-EU goods within the EU. The SAD declaration corresponding to the trip can be identified through the *containerID* and the acceptance date of the goods in the EU. Then, to calculate this indicator we are only interested in the field reporting the location of origin.

$$TSAD = (cti.stuffing.location \neq sad.location)$$

It is worth mentioning that the proposed framework is expandable and more metrics can be introduced according to the domain, the resources and the desired coverage that has to be achieved regarding the assessment of the model.

4 Experimental Results

This section presents four different studies derived by using two different CTI record computation models over two different data sources.

We have implemented two different CTI computation models: a rule-based logic deterministic model and a sequence-based probabilistic one. The rule-based model is defined by a complex set of deterministic rules manually induced by an expert (that we denoted as RULES). The second model is obtained by the application of the methodology described in [4] to construct a machine learning model based on Conditional Random Fields (that we denoted as CRF).

For the application of the proposed methodology two parameters of the CTI-QAF must be set up. For the cases presented in next subsections, we fix a maximum value of 20 % for the threshold of quantitative metrics (FDRTh), while for the qualitative indicators' threshold (QLTh) it is set up to 10 %.

4.1 CTI-QAF applied to RULES Models

In this section, CTI-QAF is applied to the CTI records obtained from two CSM data collections (Carriers 1 and 2) by the RULES models. These collections are part of the database of CSMs available within the ConTraffic project [6].

Case Study 1: Analysis of the performance of the RULES Model on Carrier 1. The input data to the CTI-QAF is the set *CTIset* computed by the RULES model. The number of containers processed for Carrier 1 by RULES was 692,233, corresponding to active containers during the period 2010–2015. The total amount of trips detected was 9,800,207.

The Stage A of the methodology measures quantitatively the performance of the *CTIset* using a random set of trips. With this aim, 72 CTI records (*CTI-Xset*) were randomly selected from *CTIset*. Then, an expert was asked to obtain the correct trips corresponding to those CTI records. A total of 70 ground-truth CTI records (*CTI-GT-Xset*) were extracted. Based on this, a False Discovery Rate (*FDR*) of 70.83 % was obtained for the *CTI-Xset* (i.e., 51 wrong trips). As the *FDR* exceed the threshold for quantitative metrics ($FDRth = 20\,\%$), the CTI-QAF recommends to reconsider the model before continue.

We compute then the *Precision*, the *Recall* and the *F1-Score* metrics for each CTI record's phase. Table 1 compiles the values of these quantitative metrics, showing problems in the proper identification of the discharging and stripping phases. Thus, the methodology provides hints in order to improve the model (e.g., creating new rules to better identify the final phases of the trips).

Case Study 2: Analysis of the performance of the RULES Model on Carrier 2. In this scenario, we give as input to CTI-QAF the set of CTI records generated by RULES (*CTIset*) from Carrier 2 data. 404,571 containers were processed (for the period 2010–2015) and a total of 4,631,965 trips were detected.

Table 1. Quantitative results for stage A of the RULES model for Carrier 1

RULES model	Precision (%)	Recall (%)	F1-Score (%)
Stuffing	100.00	88.24	93.75
First-load	98.49	91.55	94.89
Transshipment	94.12	94.12	94.12
Final-discharge	100.00	13.76	24.14
Stripping	61.11	28.21	28.60

In Stage A, the quantitative validation of *CTIset* was performed. With this aim, 90 records (*CTI-Xset*) were randomly selected from *CTIset* and an expert computed their ground-truth, obtaining 92 CTI records (*CTI-GT-Xset*).

The *FDR* for the RULES model was 14.29 % (12 wrong trips) and since it does not exceed the *FDRth* value we proceed to Stage B. The quantitative metrics computed for the different phases returned high values (all above 80 %).

In Stage B, the calculation of the qualitative indicators is carried out on the *CTIset*. Table 2 shows the results for each indicator. As they are all below the *QLTh* (10 % in our case), the assessment continues to the next Stage.

Table 2. Qualitative results on the trips obtained by the RULES model for Carrier 2

RULES model	Indicator value (%)
TLong	0.36
TShort	0.36
TMulti	0.05
TSDLong	0.73
TARLong	4.78
TLocation	0.21
TODVessel	0.09
TSAD	0.8

In Stage C, the ranked list of potentially wrong CTI records was obtained (*RListpwrongCTIset*), and its first 90 trips were selected for quantitative evaluation (*CTI-RXset*). According to the expert, 92 ground-truth CTI records were found (*CTI-GT-RXset*). The *FDR* obtained was 96.65 %, which means that the metrics for the selection of probably wrong trips have been effective.

Finally, the CTI-QAF can conclude that the model may have its most serious problems in the detection of the transshipment phase (based on the F1-Score obtained in the five phases, see Table 3).

4.2 CTI-QAF applied to CRF Models

The CTI-QAF was applied to the CTI records obtained by CRF models as well, something that demonstrates its versatility. These models were assessed using as input the same two different data sources than in previous case studies (see Sect. 4.1).

Table 3. Quantitative results for stage C of the RULES model for Carrier 2

RULES model	Precision (%)	Recall (%)	F1-Score (%)
Stuffing	50.59	94.44	65.89
First-load	65.41	65.41	65.41
Transshipment	22.22	3.95	6.70
Final-discharge	66.67	47.69	55.61
Stripping	50.00	42.86	46.15

Case Study 3: Analysis of the performance of the CRF Model on Carrier 1. In this case study, the CTIset was obtained after applying a CRF model to data from Carrier 1. 5,970,005 CTI records were computed by this model.

Stage A of the CTI-QAF measures quantitatively the performance of the *CTIset* using a random set of trips. Thus, 69 CTI records (*CTI-Xset*) were randomly selected from *CTIset* and an expert computed their ground-truth obtaining 70 CTI records (*CTI-GT-Xset*). The *FDR* was 13.18 % (i.e., 10 trips wrong), which is lower than the *FDRth* threshold (i.e., 20 %) and hence the assessment of the model can continue.

In Stage B, the qualitative indicators ranged from 0.03 % to 3.5 %, so they passed the QLth threshold which is set to 10 % for all metrics. The ranked list *RListpwrongCTIset* is then constructed, which can be analyzed by the user in order to further improve the CRF model.

Stage C allows the user to fine tuning the CTI construction model. To do this, the first 69 trips (*CTI-RXset*) from *RListpwrongCTIset* were selected and an expert calculated the ground-truth of this set, obtaining 70 CTI records (*CTI-GT-RXset*). Then, quantitative metrics for CTI records and phases are computed. The *FDR* was 95.92 %. The quantitative metrics obtained for each phase show that the stuffing phase is often wrongly computed (*F1-Score* = 34.88 %).

Case Study 4: Analysis of the performance of the CRF Model on Carrier 2. The data set (*CTIset*) contained 3,327,561 CTI records, which were computed by CRF on the data from Carrier 2 (described in previous sections).

In Stage A, we obtained the *CTI-Xset* extracting randomly 93 trips from the *CTIset*. These CTI records were given to an expert, obtaining a ground-truth set of 92 CTI records (*CTI-GT-Xset*). The *FDR* computed was 10.99 % (i.e., 12 wrong trips), which allows continuing with the Stage B.

Then, the calculation of the qualitative results was carried out, showing that in many cases the first two phases of the CTI records are not properly detected ($TSDLong$ = 7.31 %). However, as all the indicators are below the qualitative threshold, the set *RListpwrongCTIset* is obtained.

In Stage C, the first 93 trips from *RListpwrongCTIset* were selected for expert annotation. The quantitative metrics show a *FDR* of 97.13 %. A deeper analysis revealed that the model had serious problems in the detection of the transshipment (F1-Score = 16.53 %) and stuffing phases (F1-score = 23.00 %).

5 Conclusions and Future Work

In this paper, a quality assessment framework for large datasets of Container-Trips Information (CTI) is proposed (CTI-QAF). The framework combines traditional quantitative metrics with domain specific qualitative indicators in order to achieve an overall assessment of the constructed CTI records dataset. The proposed CTI-QAF is able to assess the CTI records and also highlight the aspects of the CTI computation model with more improvement potential.

Incoherent CTI records can easily jeopardize the risk analysis on the goods route. The capacity of CTI-QAF to highlight potentially wrong records can be used to support the risk assessment procedure itself by providing information on the quality of each CTI record. Moreover, knowing the potentially wrong CTI records is useful for analyzing in deep the problems of the computation model.

Two different case studies were presented to demonstrate the application of the CTI-QAF. Two different CTI record computation models have been implemented; a sequence-based probabilistic model and a decision-tree based logic deterministic one. The experimental results involved data from more than 1.1 million containers for two operators (Carrier 1 and Carrier 2), which led to the assessment of more than 23 million trips constructed by the 2 models.

The experimental results shown the effectiveness of CTI-QAF to detect models with low performance at an early stage (see Case Study 1). It is also capable to identify problematic cases, helping to improve the model (see Case Study 4).

With regard to future research lines that would extend the result of this paper, a methodology should be defined to help the users to select the appropriate thresholds to be used in the framework. Moreover, the use of composite indicators could be useful to extend CTI-QAF in two directions. First, they could be used to provide a more effective measure of the quality of a CTI record, facilitating its integration on the risk assessment process. Second, a composite indicator could be developed to provide a quality measure that could be used to directly compare different computation models between them. Finally, it is worth mentioning that the proposed framework is expandable and more qualitative metrics can be introduced depending on the domain, the resources and the desired coverage that needs to be achieved regarding the model assessment.

References

1. Adda, G., Mariani, J., Lecomte, J., Paroubek, P., Rajman, M.: The grace french part-of-speech tagging evaluation task. In: Proceedings of the First International Conference on Language Resources and Evaluation (LREC), pp. 433–441 (1998)

2. Camossi, E., Dimitrova, T., Tsois, A.: Detecting anomalous maritime container itineraries for anti-fraud and supply chain security. In: Proceedings of the 2012 European Intelligence and Security Informatics Conference, EISIC 2012, pp. 76–83. IEEE Computer Society, Washington (2012)
3. Camossi, E., Villa, P., Mazzola, L.: Semantic-based anomalous pattern discovery in moving object trajectories. CoRR, abs/1305.1946 (2013)
4. Chahuara, P., Mazzola, L., Makridis, M., Schifanella, C., Tsois, A., Pedone, M.: Inferring itineraries of containerized cargo through the application of conditional random fields. In: Proceedings of the 2014 IEEE Joint Intelligence and Security Informatics Conference, pp. 137–144. IEEE Computer Society, Washington (2014)
5. World Shipping Council: Container supply review. World Shipping Council (2011)
6. Donati, A.V., Kotsakis, E., Tsois, A., Rios, F., Zanzi, M., Varfis, A., Barbas, T., Perdigao, J.: Overview of the contraffic system. Technical report, JRC. Joint Research Centre (2007)
7. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (2002)
8. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco (2001)
9. Levi, M.D.: International Finance, 4th edn. Routledge, New York (2005)
10. Liu, T.: Learning to Rank for Information Retrieval. Springer Science & Business Media, Heidelberg (2011)
11. Manning, C.D.: Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In: Gelbukh, A.F. (ed.) CICLing 2011, Part I. LNCS, vol. 6608, pp. 171–189. Springer, Heidelberg (2011)
12. United Nations Conference on Trade and Development: Review of the Maritime Transport 2015. United Nations Publications (2015)
13. World Customs Organization: SAFE Framework of standards to secure and facilitate global trade. World Customs Organization (2015)
14. European Parliament and Council of the European Union: Comission Regulation (EC) 2286/2003 amending Regulation (EEC) No 2454/93 laying down provisions for the implementation of Council Regulation (EEC) No 2913/92 establishing the Community Customs Code5, vol. L343. Publications Office of the European Union, Luxembourg (2003)
15. Sokolova, M., Lapalme, G.: A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. **45**(4), 427–437 (2009)
16. Sutton, C., McCallum, A.: An introduction to conditional random fields. Found. Trends Mach. Learn. **4**(4), 267–373 (2012)
17. Triepels, R., Feelders, A., Daniels, H.A.M.: Uncovering document fraud in maritime freight transport based on probabilistic classification. In: Saeed, K., Homenda, W. (eds.) CISIM 2015. LNCS, vol. 9339, pp. 282–293. Springer, Heidelberg (2015)
18. Tsois, A., Cotelo Lema, J.A., Makridis, M., Checchi, E.: Using container status messages to improve targeting of high-risk cargo containers. In: Research Track at the 5th World Customs Organization Technology and Innovation Forum, Rotterdam, Netherlands (2015)
19. Villa, P., Camossi, E.: A description logic approach to discover suspicious itineraries from maritime container trajectories. In: Claramunt, C., Levashkin, S., Bertolotto, M. (eds.) GeoS 2011. LNCS, vol. 6631, pp. 182–199. Springer, Heidelberg (2011)