An Interactive Web-Based Geographical System in Assisting Location Matching Decisions

Tatyana Dimitrova, Luca Mazzola, Aris Tsois

Abstract—This paper describes an advanced prototype of a web-based geographical information system used for association of text strings describing locations with real-world geographical locations. The prototype was developed for the ConTraffic system that contains more than one billion text records describing events about cargo containers. The raw record contains a textual string describing the location where a cargo container event took place. To be able to use this raw data for analysis and visualization of the behavior of cargo containers, the location text strings need to be matched to real-world locations with known coordinates. We call this process “location cleaning”. The prototype allows the user to “clean” multiple location text strings through an interactive geographical, user-friendly and efficient web-based interface.

Index Terms—data cleaning, web-based geographical information system, interactive visualization, contextualization of geographical scraped data.

I. INTRODUCTION

Around 90% of the non-bulk cargo is transported in containers travelling by sea [1]; that is an estimated number of about 18 millions of containers per year. Because of the very large number of containers, the expensive and cumbersome procedure for inspections, only 2% of these containers [2] are physically inspected by Customs Authorities. To improve the efficiency of these controls authorities rely on risk-based analysis. In some recent studies [3]-[6] researchers tried to create new methods for analyzing itineraries of vessels and containers and provide assistance for the detection of suspicious containers. Such container route-based risk analysis has been considered in series of studies [3]-[5] as a key factor in identifying potentially suspicious consignments. This can be partially achieved through data mining processes enhanced with ontologies and semantic approaches [6]. Nevertheless, an essential part of this analysis is the visualization technique, which is widely recognized to be very powerful in anomaly detection because it takes advantage of human abilities to perceive patterns and to interpret them. Visualization can be critical in complex situations, especially when the dataset is multi-dimensional, massive and dynamic [5], [6]. When the dataset is very large, it is usually difficult to assess what kind of data analysis techniques are appropriate. In such cases, the visualization of the data could help significantly by giving an initial general view of container’s behavior. In ConTraffic [7] we developed the ConTrafficGeo application which is a web-based GIS application for geographical visualization of container traffic information [8], [9]. In certain cases, this application could indicate abnormal movements, such as movements in circles, unnecessary transshipments, long transshipment of empty container and others; such cases are found in the normal practice and may indicate situations that require more attention.

ConTrafficGeo requires that the container movement information is geographically tagged. In our case, the raw data records contain simple textual strings describing the locations where the events took place. That’s why we have “location cleaning” procedures, which try to assign a real-word “known” location to each textual description of a location. We consider as known geographical locations the locations stored in the ConTraffic DB for which we know enough details; including the location name, the country name, and the geographical coordinates. Currently the database contains more than 20,400 known locations. At present we have in the ConTraffic system more than 15 million different textual descriptions of locations. The big majority of the text strings describing locations are matched to known locations by automatic algorithms. However, a substantial part cannot be matched automatically with confidence and requires manual “cleaning”. This can be the case because they contain contradictory, irrelevant or incomplete information.

The manual location cleaning is quite a slow process. This made us to search for new solutions for multiple locations cleaning. In the current paper we present a prototype for an interactive web-based geographical information system specialized in assisting the location textual matching decision. It allows the user to search for “unclean” locations through keywords, it presents geographically relevant container traffic information and it gives the possibility to the user to simultaneously “clean” multiple location strings.

The paper is organized as follows: Section II gives an overview of the ConTraffic system and more details about the problem of location recognition. In section III we describe our solution for mitigating the problem. Finally, in the fourth section, we discuss the prototype’s advantages and limitations, achieved results, future improvements and then we conclude the paper with a discussion of the contributions of our approach.

II. ConTraffic and Location Cleaning

The ConTraffic system collects and analyses data on the movement of cargo containers. The system’s data flow and the process are represented in Fig.1. The data are gathered in
terms of little pieces of information, called “Container Status Messages” (CSM) [3]. Each CSM presents a single event or status of a container. It has a semi-structured format, describing the performed actions or the status of the containers. Currently the ConTraffic database contains more than one billion CSM records making the task of analysis or visualization quite challenging.

Each CSM includes:

1) a container identifier as assigned by the Bureau International des Containers et du Transport Intermodal [10], [11].
2) a textual description of the geographical location (called unclean location). The location described is usually a port, where the event took place. The text string may contain the name of a city, port, the name of a station/airport, the name of a country, or even detailed port or terminal codes. During the pre-processing stage, called automatic-location cleaning, a match between this string and (known) standard locations in the ConTraffic database is searched. The dataset of standard locations is built mainly using the UN/LOCODE (United Nations Code for Trade and Transport Locations) dataset, which is open source and is provided by the United Nations Economic Commission for Europe [10]. Each record contains the UN/LOCODE, a UN assigned unique (per country) code, the ISO code of the country, the name of the location, functions of the location (port, rail station, airport, etc.), geographical coordinates, etc. In case of a match, the text description is assigned to the unique identifier of the matched record. In this case we say that the text description of the location was cleaned.
3) a text description of the event occurring to the container and its load status.
4) a date and time when the event occurred.
5) additional details, if they are available; for example vessel name, voyage identity number, etc.

Since the CSMs are text messages gathered from multiple and heterogeneous data sources, the data can be imprecise and ambiguous [12]. This creates many difficulties in interpreting and integrating the data. In many cases, the data are not complete and it may even contain errors. Moreover, the data are semi-structured, which means that the information is not following any standardized data format. These problems create quite big difficulties in the interpretation of the location text descriptions. We rely on some algorithms for automated location cleaning, realized in PLSQL, but they cannot cover all the cases, so manual intervention is constantly needed. Many of the textual descriptions don’t contain all necessary information to allow for an automatic match (for example, they contain only the name of the country or only the name of the location, but very often a location with such a name exists in more than one places in the world). In many other cases the textual description is inconsistent (for example, you may be given the name of the location and the name of the country, but according to our DB there is a location with such name in another country, not in the country which was reported in the textual description).

To assist the location cleaning procedures we have collected and stored in our DB detailed information about known locations. This information we collected mainly from UN/LOCODE dataset. Since there are gaps in this datasets we used also another open source geographical data sets [13]. For verification and modification of the collected data in ConTraffic DB we developed a web-based application GeoLocSI (Fig.2) [14] which has the target to improve the accuracy of the geographical data and the precision of the geographical coordinates of the known locations.

![Fig. 1. Overview of the ConTraffic system.](image1)

The application allows the user to select a country and review all geographical locations in the selected country in textual and geographical way. The textual presentation is organized in an interactive table, allowing extra details for each row, selecting of single location and filtering the data. The graphical presentation allows the user to review the geographical position of a single location, part of the
locations or all locations in the selected country on the map. For each single location, the user can modify the textual data in the right panel. The geographical coordinates can be modified also by dragging the icon to the correct geographical position on the map. The application is realized by using PHP, JavaScript, SQL, HTML, Datatable Plug-In of JQuery and Google Maps API v.3.

To assist the manual location cleaning procedures ConTraffic uses a web-based application which allows the user to perform location cleaning of a single location. It is developed by Java. During its exploitation it was established that many locations follow similar model and an additional possibility to clean multiple locations could speed up the process of location cleaning. Another limitation of the used application was that there was not available additional information about the container itineraries where the unclean location was used. The location cleaning was performed based only on the textual information and there was not a possibility for any geographical visualization. There was not possibility for user interactivity, like user request for searching patterns of unclean locations, review of multiple locations at the same moment or client-side interactivity with the user defined parameters.

These limitations made us to create a new prototype giving more freedom and interactivity with the user during the location cleaning. To develop it we used a similar architecture of GeoLocSI’s architecture, and the prototype is realized as an interactive web-based geographical system which assists the multiple locations textual matching decision. The design and the implementation of this prototype are described in the next section.

III. REALIZATION OF AN INTERACTIVE WEB-BASED GEOGRAPHICAL SYSTEM IN ASSISTING LOCATION TEXTUAL MATCHING DECISION

For the development of our prototype for Multiple Location Cleaning (MLC) we used the following programming languages: PHP (for the dynamic part of the webpages), HTML (for the static part of the webpages), JavaScript (for realization of the user interactivity), SQL (for creation and execution of request to the Oracle DB), Google Maps API (for mapping and handling the mouse events) and Data Table Plug-In for JQuery library (for tabular presentation of the data). A scheme of the MLC architecture can be seen in Fig.3. The collected set of known locations with geographical coordinates is stored in a table “GIS_LOCATIONS” in our DB.

The main source of the application is in “MLC_base.php” file. It is responsible for initial loading of the application and visualization of any user requested information. This file uses some assistance files – “dbinfo.inc.php” (where all necessary data for connection to the DB such as username, password, sid, address, etc. are stored), all_unclean_locations.php (a php program which allows the user to review all unclean locations in ConTraffic DB in an interactive Data Table), all_unclean_locations_Carrier.php (a php program which allows the user to review all unclean locations for selected by the user carrier), unclean_location_details.php (a php program which allows the user to review all cases of usage of certain unclean location with detailed traffic information, such as the previous and next locations from the containers itineraries where the unclean location was used) and sql_execute.php (a php program which is responsible for the execution of SQL update requests). The main screen of the prototype can be seen in Fig.4.

![Fig. 3. MLC Architecture.](image)

From the main page of the MLC application the user can make immediately a request to review all unclean locations in the DB (Fig.5). The result is visualized, inside another tab-window, in an interactive table where the data can be filtered by keywords and sorted according to each of the visualized parameters (the data in Fig.5 is sorted in alphabetical order according to the location name). Using these two functions it is quite easy to recognize the location patterns used for presenting the same location. In Fig.6 it is possible to see the advantage of the filtering - 21 unclean locations were found just by filtering out of the initial set of 26,745 locations.
The practice shows that sometimes, it is more useful if the research of unclean locations is performed for only one data provider at a time. A button on the MLC application allows the user to filter and sort the unclean locations only for a specific carrier selected by the user (Fig.7). In this case, it is quite easy to see the patterns used from selected carrier for presenting the same location (in Fig.7 for example it is possible to see an example of the variations for textual description of the international port of SINGAPORE used by carrier RCL).

After having an idea what patterns the carriers use for describing the location, the user can refer to the main page of MLC and perform multiple locations cleaning. From the list box in the right side, the reference location must be selected (the listing includes all “known” locations for ConTraffic stored in GIS_LOCATION table). After the selection of a reference location, the geographical position of this location will be visualized by an red icon over a Google map and all the detailed information for this location will be presented in text fields (Fig.8).
Once the user inserts the pattern for search, a button “Search” recalls the MLC application with the proper user input parameters and the PHP script connects to ConTraffic DB and extracts the unclean locations corresponding to the requested user pattern. Then, the location cleaning goes through a few steps. The results from each step are visualized in tabs where the user can perform the necessary actions.

The first step is the visualization of all found unclean locations which follow the user pattern (Fig.9). The listing is visualized in an interactive Data Table where the user can sort and filter the information. From this table, the user selects multiple locations which properly match to the selected reference and only these locations will be considered in the next steps for multiple locations cleaning.

Fig.9. Listing of all found unclean locations according the user pattern.

In the next step the user can review only the selected locations and must confirm the selection in order to continue ahead with the multiple locations cleaning (Fig.10).

Fig.10. Listing of the selected from the user unclean locations.

In the step after that, the proper SQL requests are performed (according to the selected reference location and the selected unclean locations) and visualized. Then the user can decide to execute the SQL request through the MLC or copy the queries for further modifications and execution directly in the DB interface (Fig.11).

Since often the textual description is not enough even for a human to interpret it correctly and to recognize the known location, there is a need of additional information which could assist the decision maker. In some other experiments [15], [16] we are exploring an approach that tries to use more dimensions of the container traffic data together with the geographical data in order to provide a better assistance. Unfortunately, the results are not yet satisfactory and more work and validation is necessary. However, the results show us that the approach is promising, so in the proposed prototype we are offering an assistance program which provides some more detailed information about segments of the container itinerary where the unclean location is included. This information is accessible by an assistance PHP program (unclean_location_details.php) (Fig.12). To work with this application the user needs to know the identification number of the unclean location (visible in the listing of the unclean locations visualized by the MLC application - Fig.9). This identification number is used as an input parameter. All unclean locations in the MLC base application are presented by their location unique codes which are hyperlinks. The user can use these hyperlinks to call the extension application and review the container traffic information related to the unclean location. Every time when the application is called with input parameter, it searches in our DB for all cases where the textual description of the unclean location is used and presents detailed information for the itineraries where it was used in two ways: a textual presentation in an interactive table and a graphical view, enhanced by geographical map and icons. The textual presentation includes the following information: the container identification number, the short name of the carrier, the name of the previous location (PL), the date of the last event on the PL, the geographical coordinates of the PL, the date of the event on the unclean location, the name of the next location (NL), the date of the last event on the NL, the geographical coordinates of the NL, the durations of the travelling between the PL and the unclean location, the duration of the traveling between the unclean location and the NL.
The geographical presentation shows pinpoint indicating the positions of the PL and NL (in case these are defined and cleaned) for selected case by the user and the approximate area where the unclean location could be by circles. The first case is selected and loaded by default. According to the Ref.[17] for economic reasons “the optimum vessel for deep sea container trading is typically operated at 21-22 knots”.

That’s why in our application we consider 22 knots for default speed for the cargo container (which could be changed in any time). Considering also the fact that the container travels not only by the sea but also on land and in this case the typical speed will be quite different, the user can change the speed to speed more suitable to the typical speed for traveling on land.

The areas of the possible position of the unclean location are presented by circles with centers (the PL and the NL – respectively the red and blue icon in Fig.12) and radiuses – the distances $D_1$, $D_2$. These distances are calculated as follows:

\[ D_1 = V \times T_1 \]  
\[ D_2 = V \times T_2 \]

where $V$ is the speed m/s (changing this speed the user can review the different calculated zones), $T_1$ is the duration of the travelling between the PL and the unclean location in seconds and $T_2$ is the duration of the travelling between the unclean location and the NL. We would like to mention that this visualization is just indicative and it presents approximate zones of reachability without considering many facts, such as logistics of the container itinerary, ship travel routes are not following straight lines and the ships don’t travel with a constant speed, sometimes they stop for fixing, or unplanned reasons, the container itinerary is uncompleted, because it contains only active locations (location where some event occurs, but many times the container travels to ports where no event occurs for it), there are some errors in the data and other.

In the specific example in Fig.12, there are only 2 cases and the visualization of the first case could help us to restrict the area of possible position of the unclean location and confirm our suggestion for the port of Singapore.

Another example of unclean location is presented in Fig.13. In the raw textual description is included only the name of the location “Malton” which is not enough for performing location cleaning. As it can be seen using the information related to the previous and next locations, the application allows restriction of the reachable zone. From the visualization the user could recognize four zones – everything outside the colored area, reachable zone according to the previous location, reachable zone according to the next location and the intersection of the reachable zones. Further option of the application will allow the user to see listing of all known locations in the restricted zones which will help him more in the matching process.

Fig. 14 presents a case in which the travelling days from previous location to the unclean one is null, indicating that date of departure from PL and date of arrival on the unclean location are the same day. This can be interpreted as a very small or null geographical distance between the two considered locations (as it is in the presented case) or as an error in the data.
Unfortunately, sometimes there are some really difficult cases for which our application cannot help much. Examples of these cases can be seen in the listing presented in Fig.15 and the geographical visualization in Fig.16. For some cases (N13-20), the PLs and/or the NLs are unclean still locations, so their geographical coordinates are still unknown and the calculation of restricted zone is not possible since it is not known the centre of the reachable zones.

![Fig.15](image1.png)

**Fig.15.** Difficult cases for solving

In other cases (case N16, 18, 20) the duration of the travelling is too long, so practically the circle of reachable zone covers most of the Earth and the position of the unclean location could be anywhere. We consider duration of travelling longer than the threshold of 8 days as not indicative parameter and highlight the value with a red background (as can be seen in Fig. 16). In some further research it could be used also for indicator of anomalies and suspicious itineraries. In most of the cases, obviously the reason could be one of the mentioned facts or just gaps in the DB or data errors.

![Fig.16](image2.png)

**Fig.16.** Presentation of a case where the number of traveling days is higher than the threshold and the reachability area covers the entire Earth.

The tool is actually used for supporting the location cleaning procedures of the CSM collected in the ConTraffic project. Despite the short time from which it has substituted the original Java application, the experts involved in the task reported positive feedbacks on its adoptions, even though some suggestion to simplify the interaction patterns and some indication to improve the interface are being collected from them. An interesting information emerged from the statistical analysis of the location cleaned with this Multiple Location Cleaning tool: the relevance of the records cleaned - for the interpretation of the global CSM positioning - is higher for this MLC tool than for any pre-existing adopted approach, thanks to the possibility to consider the pattern-matching string in a aggregated way, making emerge the ones that are globally more relevant.

**IV. DISCUSSION AND CONCLUSION**

In this study, an interactive web-based geographical system has been presented. The prototype was developed to assist the location cleaning process, the detection of common patterns and review the additional information for certain location textual descriptions. The prototype allows for a fast interactive review of text strings describing locations supporting the location matching decisions in various ways. The advantage of the tool lies in its ability to “clean” simultaneously multiple text location strings which are similar and match a given keyword or pattern. Indeed, in practice, very often we have hundreds (or even thousands) of similar text strings that represent the same location. These strings are similar because they can be identified by a pattern or a keyword but they are not equal; possibly because they contain in addition to the pure location description some numerical identifiers or other characters. The tool allows for the simultaneous treatment and “cleaning” of all these strings making the location cleaning process much more efficient. The expert also has the possibility to review a segment of container itineraries which include the unclean location and restrict the geographical area to a reachable zone, that present a higher probability of including the unknown location. The geographical zones are calculated according to the container travelling information (time of travelling, calculated between the two sequence locations) and the user selected speed for traveling. The tool seems to be promising for the task it was designed for and we are now trying to improve it, based on the indications and feedbacks received from the users.

Unfortunately the proposed prototype has also some limitations due to problems in the data on which it relies. They come from some facts:

- sometimes the unclean locations are surrounded by other unclean locations in the same itinerary;
- the data include contradictions or discrepancies (result from typing and semantic errors);
- in some cases the same locations or countries are described in different ways or language;
- presence of special letters (not included in the Latin alphabet);
- gaps in the container itinerary (identified by the
appearance of too big time intervals between the sequence locations;  
- the unclean location is not surrounded by other locations (start or end of the trip);  
- the travelling speed can vary significantly (this is directly reflected into the calculated zones of reachability).

The prototype has been successfully integrated in the ConTraffic system and it is currently used for multiple locations cleaning. During its usage, the users were able to detect some common patterns and clean bigger number of unclean locations. The new functions of the proposed prototype are the possibility for multiple locations cleaning and available additional information about the container itineraries including the unclean location.

Many future improvements of the prototype can be considered. Some of them can be: filtering out the cases not having previous and next clean locations, filtering out the cases having a travelling time bigger than the threshold, presenting the cases already cleaned with similar textual description, presenting some statistical data calculated from the entire container traffic data (for example: most frequently used next/previous location to the locations included in the segment of the container itinerary, typical interval of time for travelling to them, the typical trajectories of the services by a specific carrier, vessel voyages and itineraries, listing of known locations in the defined reachable zones).

REFERENCES


Tatiana V. Dimitrova was born in Bulgaria, Ruse, 12th of October, 1979. In 2002 she got her Bachelor degree in Computer Systems and Technologies from the Technical University of Ruse, Bulgaria. After that, she continues her university education and in 2003 she got her Master degree in Computer Technologies. During the university study, she followed some additional courses for additional professional qualifications. In next few years, she worked on PhD thesis which was connected with brute image analysis and development of a system for brute age determination. During 2006 she was at student specialization for 6 months by “Erasmus” program in Computer Science Department, Cyprus University. In 2009 she got her PhD Degree in Bulgaria. Later the same year, she started like stagier in Seal Identification Laboratory, IPSC, JRC, Italy where she worked on the development of a system for automatic verification of copper-brass seal’s images. Since August, 2010, she is a grant holder in Maritime Affairs Unit, IPSC, JRC and works on data analysis and is responsible for developing advanced interactive web-based geographical applications for visualizing large scale dynamic data for container movements for analyze the flow of goods. This includes the identification of coherent container movements, monitoring of the port traffic and identification of anomalous movements. Her professional interests include image processing, web development, geographical information systems and geographical visualization.

Luca Mazzola was born in Bergamo, Italy, March 31st, 1979. He received his Master degree in Engineering in Computer Sciences from Politecnico di Milano, Italy in 2004. After a period working at CILEA (now CINECA), one of the three Italian inter-university consortiums for ICT in Segrate (Mi), Italy, he moved to “Università della Svizzera italiana” (USI) in Lugano, CH. Here he participated in different research projects and service conception, covering different areas (from eLearning to folksonomies data reuse passing through knowledge extraction for huge data sets). Recently, he worked also as researcher at Politecnico di Milano in the field of Decision Support System and e-Health. He is currently grant holder at IPSC, JRC and contract professor of Medical Informatics at DEIB - Politecnico di Milano. His research interest spans from secondary usage of data to knowledge engineering, from ontologies to application of IT to security and health.

Aris Tsois studied Electric and Computer Engineering at the National Technical University of Athens, Greece from which he obtained his diploma in 1995. He received from the same university his PhD in data bases and data warehouses in 2005. He worked at the National Technical University in several research projects for more than 10 years as well as IT consultant, software architect and IT developer for private companies. In 2005 he joined the Joint Research Centre of the European Commission as a scientific project officer where he is currently managing the ConTraffic project. Mr. Tsois is affiliated with ACM since 1995 and has published book chapters as well as several papers in journals, workshops and international conferences including the VLDB. His research interests include database systems, query optimization, conceptual data modeling and artificial intelligence.