Resolution of Geographical String Name

Luca Mazzola[1]*, Aris Tsois[1], Tatyana Dimitrova[1], Elena Camossi[1], Alberto V. Donati[2], and Mauro Pedone[3]

[1] European Commission, Joint Research Center *(JRC)*
Via Enrico Fermi 2749, Ispra, Varese, Italy – I-21027
Email: {luca.mazzola, aris.tsois, tatyana.dimitrova}@jrc.ec.europa.eu,
elena.camossi@gmail.com
[2] Polimedia – Email: albertov.donati@gmail.com
[3] SERCO – Email: pedone.mauro@gmail.com

Abstract. This paper presents a model, and a first partial implementation, for location resolution of string description in the status messages generated by maritime container traffic. The work was carried out in the context of the *ConTraffic* project, which deals with container traffic data. The model is based on the usage of different data dimensions, such as string similarity, trajectories similarity and most frequent patterns. The realized interface integrates these distinct dimensions, and uses them to provide a map-based view to support a human expert in associating the string description provided in the raw record to a location.

1 Introduction

In the domain of mobility data management and exploration, the use of multiple and heterogeneous data sources is quite common. When a system uses data created by different systems, with different semantics and different data models the need for data interpretation and integration is evident in order to produce an unified and valuable result for the purpose of a given task [1].

The *ConTraffic* pilot project running at the Joint Research Centre of the European Commission is one such system based on many data sources that need to be transformed and integrated. ConTraffic deals with information on the status and movement of cargo containers. The aim of this pilot project is to develop new techniques and processes to help authorities in their effort to control the flow of cargo containers. It is estimated that today more than 20 million cargo containers are used worldwide to transport about 90% of the world's cargo. The key concept in ConTraffic is the usage of available data sources, currently not exploited by authorities, and the application of state of the art analysis techniques on the collected mobility data [2].

The basic information unit in ConTraffic is the *Container Status Message* (CSM). Each CSM describes a logistic event of a particular cargo container, such as the loading of a container on a vessel or the delivery of a full container

* Corresponding author: luca.mazzola@jrc.ec.europa.eu - mazzola.luca@gmail.com

to its final destination. Such events can be represented by an ontology [3] and can be analized using a semantic approach to extract valuable relations [4]. Each CSM contains the date (and sometimes a precise time) and a textual description of the location where the event took place, besides other fields.

In order to be able to use the CSMs in various analysis processes, and in order to be able to integrate the data collected from the different sources, the textual location description in each CSM must be mapped to a real-world location. In Contraffic we use the *UN/LOCODE* [5] taxonomy, that is a set of locations for trade and transport purposes collected by the United Nations, as a reference table of real-world locations.

In this paper we address the issue of resolution of strings that represent locations of containers. The method goes beyond string similarity and uses the trajectories of containers and the fact that many containers follow the same routes. In ConTraffic, based on the hundreds millions of CSM records available, one can reconstruct millions of trajectories that represent the movement of containers. This is possible by using the location description and the date/time of each CSM record. However, in this domain there is a significant uncertainty about the exact trajectories as the CSM records do not describe precisely when the movement of the container actually takes place, how the container moves from one location to the other or for how long the container remains at each location. Furthermore, the different means of transport used, like vessels, trains or trucks, have different speeds, complicating even more the estimation of the real trajectory.

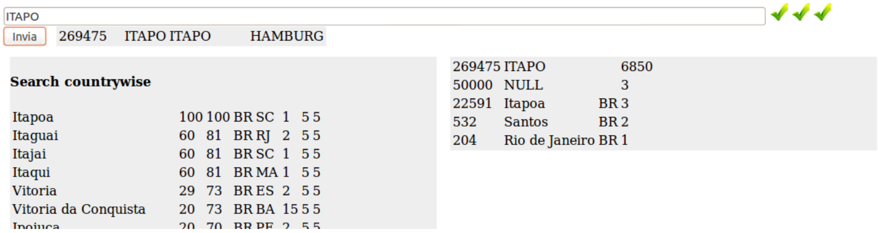
In the next section we define the specific problem we deal with and in the third section we present our analysis and we propose a possible solution for it.

2 Problem Definition

The issue we try to address is when a text string identifying a location cannot be matched with confidence using string similarity to any particular location in the reference list. The reasons that lead to this lack of classification can be grouped in three categories:

1. the string can be matched to one location only, but the confidence is not strong enough (for example,only a subpart of the string matches).
2. the string can be potentially associated with more than one location, and none of them with enough strong confidence.
3. no match with any location in the reference list is found.

The above cases can occur whenever the matching process must respect a “confidence” level regardless of how that level is defined. This means that we want to be sure enough (the threshold over the confidence level) about the association we are proposing. For the specific purposes persecuted, in ConTraffic we try to avoid any misclassification of location strings so the rules that define confidence are quite strict (mostly exact string matches, maybe enhanced with



ITAPO ✓✓✓

269475 ITAPO ITAPO HAMBURG

Search countrywise

Itapoa	100	100	BR	SC	1	5	5
Itaguai	60	81	BR	RJ	2	5	5
Itajai	60	81	BR	SC	1	5	5
Itaqui	60	81	BR	MA	1	5	5
Vitoria	29	73	BR	ES	2	5	5
Vitoria da Conquista	20	73	BR	BA	15	5	5
Ipojuca	20	70	BR	PE	2	5	5

269475	ITAPO		6850
50000	NULL		3
22591	Itapoa	BR	3
532	Santos	BR	2
204	Rio de Janeiro	BR	1

Fig. 1. The first two dimensions, as represent textually for the location string "*ITAPO*"

permutations). In these cases, the human expert intervention is required to resolve the string and to associate it with one of the entries in the reference list, if possible. Sometimes this cleaning process remains unfeasible even for the human experts.

The concept of the moving point [6] is well-known in the maritime transportation domain, and it can be seen as the abstraction of the movements of a container.

The concrete problem that we address in this paper is how trajectory information of large sets of similar objects can be used to resolve an unknown intermediate location in a trajectory of one such object.

In order to consider the problem in its broader and abstract form we give the following problem definition:

Definition 1. *Assume R is a set of trajectories, S is the set of moving objects mentioned in R and consider that each subsegment of a trajectory is called "trajectory segment". Assume also that each starting or ending point of each trajectory segment in R is considered a location and it is assigned a location name (not necessarily unique) and a unique geographical coordinate pair. Let $L = \{l_1, l_2, .., l_k\}$ be the set of all locations $l_i = (c_i, d_i)$ in R, where $c_i = (latr_i, lonr_i)$ is the couple of value representing the coordinates and d_i is the textual string encoding the name.*
Given that x is the text string describing the location where the moving objects $M = m_1, m_2, .., m_j$, with m_i in S were at the corresponding time intervals $T = t_1, t_2, .., t_j$, how can one use R, M, T and L to map x to one of the members of L with the required confidence level?

3 Analysis

To address the problem we will decompose it in smaller independent sub-problems based on the information used:

[*A* - STRING dimension] Use the locations names $d_1, d_2, .., d_k$ in L to select a number of candidate locations matching x based on text string processing techniques. In general, for each location l_i in L one can calculate a confidence level for the mapping $x \Longrightarrow l_i$ and then use a confidence level threshold to select which locations (none, one or many) are the final candidates for x. We can apply

different algorithms to select the confidence level; exact string match, sub-string match, string distance functions are typical examples [7, 8], but also other are used [9]. They can be used alone or in combination with a semantic analysis of the text string x, removing from x redundant words, characters or identifying the target country (or region of the country) where the location seems to be. In the best case this technique returns exactly one candidate.
For example, if x is the string "Terminal A at port of Napoli" and in L we expect that there will be only one location matching this description (namely "Napoli") then a sub-string matching technique would return only this as candidate location to map to. Unfortunately locations exist, in France and in Switzerland, whose name is "Port", generating a multiple matching that needs to be solved.

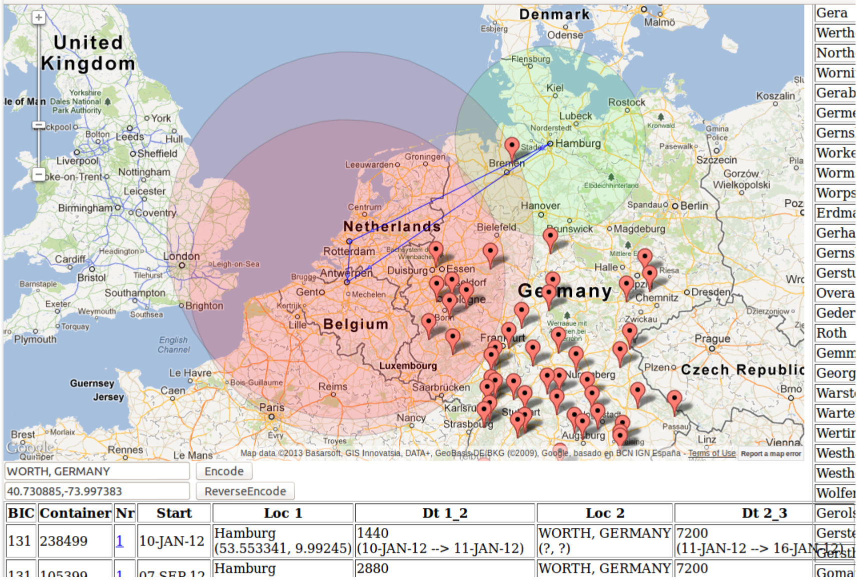
[B - RELEVANT TRAJECTORY dimension] Use the relevant[2] trajectories of the objects in M, in combination with the given time intervals T and the geographical coordinates of the locations in L to predict a set of candidate matching locations that can be associated to x. The prediction algorithm in the first phase should find candidate locations for each m_i in M separately and then combine the results to produce the final set of candidate locations (for example, using a set intersection function). Note that if the quality of data is not guaranteed (like in the case of ConTraffic), then combining the results of the individual m_i may not be straightforward (exclusion of outliers may be needed or probabilistic fuzzy techniques). There are several techniques to predict grid-based candidate locations [10]. When choosing this technique, one must carefully understand the problem domain. For example, in ConTraffic any prediction based on linear interpolation would fail because containers do not maintain any kind of constant speed or direction when moving from one location mentioned by a CSM to the next.

[C - TRAJECTORY representatives] The trajectories in R are used to *predict* the location x. Typically this can be done by clustering the trajectories to reveal the common patterns of movement (finding a representative trajectory). Then, the clusters that contain a relevant trajectory of an object in M can be used to predict candidate locations for x. Just as in the previous case the independent results achieved for each m_i in M can be combined to reduce the set of candidate matching locations. This technique can be used in domains where many objects follow similar trajectories (or at least similar routes). In general the results are expected to depend a lot on the clustering parameters used and on the topology of the set of trajectories R. Overall, if more than one of the above techniques is used, then the obtained results can be combined to improve the results by reducing the set of candidate matching locations.

3.1 Implementation

Based on the model presented we implemented a software tool to help experts associate the strings with the relative locations. It partially implements the model

[2] For each m_i and its associated time interval $t_i = (start_i, end_i)$ at x there can be at most 2 relevant trajectories of m_i in R: the trajectory that ends just before $start_i$ and the trajectory that starts just after end_i.


United Kingdom
Denmark
Netherlands
Belgium
Germany
Czech Republic
Luxembourg
Hamburg
Bremen
Berlin
Paris
London
Map data ©2013 Basarsoft, GIS Innovatsia, DATA+, GeoBasis-DE/BKG (©2009), Google, basado en BCN IGN España - Terms of Use
Report a map error

WORTH, GERMANY
Encode

40.730885,-73.997383
ReverseEncode

BIC	Container	Nr	Start	Loc 1	Dt 1_2	Loc 2	Dt 2_3
131	238499	1	10-JAN-12	Hamburg (53.553341, 9.99245)	1440 (10-JAN-12 --> 11-JAN-12)	WORTH, GERMANY (?, ?)	7200 (11-JAN-12 --> 16-JAN-12)
131	105399	1	07-SEP-12	Hamburg	2880	WORTH, GERMANY	7200

Gera
Werth
North
Worni
Gerab
Germe
Gerns
Worke
Worm
Worps
Erdma
Gerha
Gerns
Gerstu
Overa
Geder
Roth
Gemm
Georg
Warst
Warte
Wertin
Westh
Westh
Wolfer
Gerols
Gerste
Gersth
Goma

Fig. 2. The Application developed: an uncleaned location '*WORTH, GERMANY*'

presented above– namely the first and partially the second and third approaches– and uses the computed sets of candidate locations to provide the information to expert in two formats: textual lists (such the ones in the Fig. 1) and as graphical elements into a map, as presented in Fig. 2.

The two columns in Fig. 1 represent respectively the candidate locations in the string distance dimension, and the most frequent location of the trajectories that share the previous and the next location string with the current analyzed case x. In Fig. 2 instead are presented the actual interface for the final user, in which the information is filtered and restricted to a limited number in order to be useful, when presented on a map. For the *string* dimension – based on a metric simile to the Jaro-Winkler[11] one–, the search is bounded to the nation included in the raw string, if available, and limited to the first 20 most simile candidate locations in the set. The information is then included in the map as red pinpoints.

In the *relevant trajectories* dimension, the ranking is based on frequency of locations co-occurrence with the same previous and next steps. It is later limited on the temporal duration of the segment. Here a problem emerges for the presence of sometimes incomplete timestamps (just composed by the date part). This dimension is represented in the map as circles, with the centers in the known location and the radius proportional to the time distance of the events[4].

[4] Defining the previous and current location respectively as PL and x –with the associated timestamps T_{PL} and T_x– the radius of the circle R_{PL} centered in PL is

For every trajectory, the previous location is depicted with a green circle, while the red ones represents the next location.

An initial tentative to meaningful use the information related to the third sub-problem (use of trajectories to predict the location) was done, as can be seen in the lower part of the Fig. 2, by showing the 10 shortest sub-trajectories having in common the previous and next locations (with regards to x and t_i for a given m_i). We choose to represent it as straight lines connecting the previous and next location, to give an idea of the trajectory, if the unclean location was not present or not cleanable.

References

1. Mazzola, L., Eynard, D., Mazza, R. (2010). GVIS: a framework for graphical mashups of heterogeneous sources to support data interpretation. HSI2010. 3rd IEEE Conference on Human System Interactions, 2010. pp. 578 - 584.
2. Varfis, A., Kotsakis, E., Tsois, A., Donati, A. V., Sjachyn, M., Camossi, E., Villa, P., Dimitrova, T., and M. Pellissier. (2011) "ConTraffic: Maritime container traffic anomaly detection." In MAD 2011 Workshop Proceedings, p. 113.
3. Villa, P. and Camossi, E. (2011). A description logic approach to discover suspicious itineraries from maritime container trajectories. In GeoSpatial Semantics (pp. 182-199). Springer Berlin Heidelberg.
4. Camossi, E., Villa, P., and Mazzola, L. (2013). Semantic-based Anomalous Pattern Discovery in Moving Object Trajectories. arXiv preprint arXiv:1305.1946.
5. United Nations Code for Trade and Transport Locations, Last accessed on Apr. 12th 2013. Available at http://www.unece.org/cefact/locode/welcome.html
6. Güting, R.H., Böhlen, M. H., Erwig, M., Jensen, C. S., Lorentzos, N. A., Schneider, M., and Vazirgiannis, M. (2000). A foundation for representing and querying moving objects. ACM Transactions on Database Systems (TODS), 25(1), 1-42.
7. Sehgal, V., Getoor, L., and Viechnicki, P. D. (2006). Entity resolution in geospatial data integration. In Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems (pp. 83-90). ACM.
8. Kang, H., Sehgal, V., and Getoor, L. (2007). GeoDDupe: a novel interface for interactive entity resolution in geospatial data. In Information Visualization, 2007. IV'07. 11th International Conference (pp. 489-496). IEEE.
9. Martins, B., Anastácio, I., and Calado, P. (2010). A machine learning approach for resolving place references in text. In Geospatial Thinking (pp. 221-236). Springer Berlin Heidelberg.
10. Krumm, J.,and Horvitz, E. (2006). Predestination: Inferring destinations from partial trajectories. In UbiComp 2006: Ubiquitous Computing (pp. 243-260). Springer Berlin Heidelberg.
11. Winkler, W. E. (2006). Overview of Record Linkage and Current Research Directions. Research Report Series, RRS.

proportional to the time difference of the events $R_{PL} \propto (T_x - T_{PL})$. The same applies for the circle centered on the next location NL.