Georeferencing Facts in Road Networks

Fábio da Costa Albuquerque[1,3], Ivanildo Barbosa[1,2], Marco Antonio Casanova[1,3], Marcelo Tílio Monteiro de Carvalho[3]

[1]Departament of Informatics – PUC-Rio
Rio de Janeiro – Brazil

[2]Department of Surveying Engineering – Military Institute of Engineering
Rio de Janeiro – Brazil

[3]TecGraf – PUC-Rio
Rio de Janeiro – Brazil

{falbuquerque, ibarbosa, casanova}@inf.puc-rio.br,
tilio@tecgraf.puc-rio.br

Abstract. *Information about a location can be imprecise and context-dependent. This is especially true for road networks, where some streets are long or two-way, and just the name of a street may represent low-value information for certain applications. To improve precision, geocoding commonly includes the number of a building on a street, the highway location, often indicated in kilometers, or the postal code in a town or city. One can also improve the description of a location using spatial attributes, because they are familiar concepts for humans. This article outlines a model to precisely georeference locations, using geocoding and routing services and considering the natural attributes used by humans regarding locations.*

1. Introduction

In this article, we address the problem of inferring the location of facts that affect road conditions by analyzing real-time data retrieved from dynamic data sources on the Web. In general, the location of such facts is useful for real-time applications that monitor moving objects and that support decision making. For example, car crashes and road blocks are relevant to such applications because they affect the traffic flow by reducing the average speed and imposing changes on the planned route. However, to be useful, the location of such facts must be estimated as accurately as possible. Furthermore, they must be provided as timely as possible, which justifies exploring dynamic data sources on the Web.

The most common way to georeference locations is to use geocoding techniques, which can be defined as a process to estimate the most accurate location for a set of geographic points from locational data such as postal code, street name, building name, neighborhood, etc. As summarized by Goldberg, Wilson and Knoblock (2007), geocoded data that used to cost $4.50 per 1,000 records as recently as the mid-1980s, quickly moved to $1.00 by 2003, and can now be done for free, using online services, which however may have limitations, such as the maximum number of requests per day. For example, Yahoo! PlaceFinder allows up to 50,000 requests per day, while Google

allows 2,500 requests, Bing allows 15,000 requests, and CloudMade provides unlimited access to this service completely for free.

Information about a location can be imprecise and context-dependent. In a road network, where streets may be long or two-way, just the name of a street may represent low-valued information for certain applications. To improve precision, geocoding commonly includes the number of a building on the street, the highway location, often indicated in kilometers, or the postal code. Another way to reference locations, frequently used in human communication, is to use a proximity attribute, declaring that location *A is near* location *B*, rather than directly using the address of location *A*. Another relevant aspect of location description using natural language is the direction attribute, i.e., the direction of a street toward a location.

In this article, we outline a model to georeference the location of facts, using geocoding and routing services, from spatial descriptions commonly found in human communication. To validate the model, we describe a prototype application that uses structured traffic-related news in natural language to infer locations. The prototype application is part of a larger system to monitor moving objects in an urban environment [Albuquerque et al., 2012].

The article is organized as follows. Section 2 describes our motivation. Section 3 introduces the geocoding model. Section 4 presents the prototype application. Finally, section 5 draws some conclusions.

2. Motivation

To motivate the discussion, consider the following scenario. Every day, Twitter text messages ("tweets") with traffic-related contents are published by institutional or individual users as a collaborative initiative. Institutional tweets, such as those published by CET-RIO[1], are fairly well-structured and can be used as raw input data in the context of our target application. Each traffic-related tweet contains one or more simple facts (such as traffic intensity) or describes events (such as accidents or road blocks) and their respective location. We do not distinguish between simple facts and events here, and refer to both as facts. Retrieving these associations from raw text is not trivial because there is no commonly expected format or template for natural speech. In order to associate the facts to their accurate locations, we use a traffic-related fact structure, as explained in Section 4.1.

The naive use of location as input to the georeferencing process may produce imprecise results. As an example, consider the text illustrated in Figure 1: "Car accident on street *A*, located at district *K*, in the direction of Hospital *X*, near street *B*". Suppose also that street *A* is a two-way street and is 5 kilometers long.

If the geocoding process outputs only "*street A"*, then the information will be quite inaccurate: we do not know the exact location of the accident along the 5 kilometers, or in which street direction it occurred. On the other hand, a geocoding service that qualifies *"street A"* with *"near street B"* provides valuable information that

[1] http://twitter.com/CETRIO_ONLINE

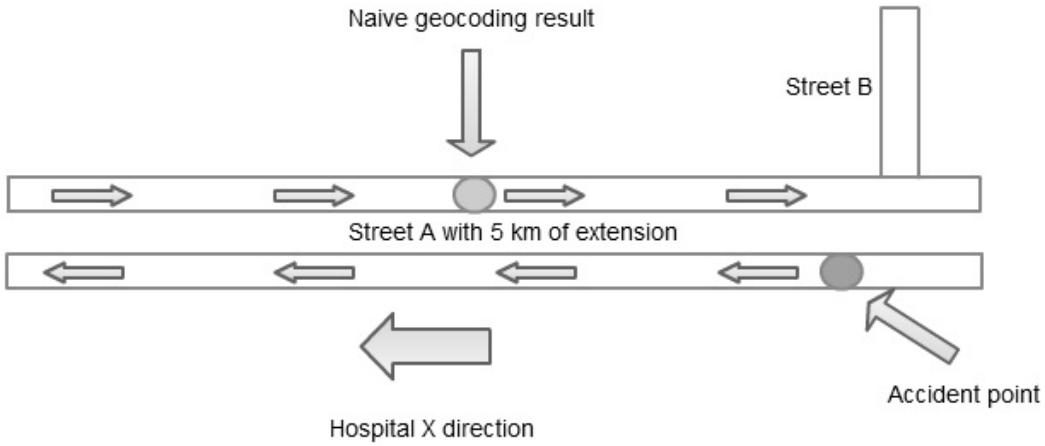

Figure 1. Example of naive geocoding.

can be used to narrow the location of the accident, whereas the text fragment *"in the direction of Hospital X"* indicates which street direction was affected.

The use of additional predicates, based on spatial references, also helps improving the description of a location. In the above example, it is easier for a driver to identify *Hospital X* along a street than to check the number of the buildings. Once the hospital location is known, spatial reasoning will provide additional information. Therefore, references like *near*, *intersecting*, and *located at*, although not deterministic, narrow the scope of the location-based analysis.

3. Geocoding Model

This section presents the geocoding model and how it is used to increase georeferencing precision, relying on geocoding and routing services available on the Internet.

3.1. A Brief Outline of the Model

As discussed in Section 2, we typically use additional data to improve the precision of a location of interest. The model we adopt, summarized in Figure 2, has the following entity sets and relationships (we indicate only the most relevant attributes for brevity):

Entity Sets

Fact the set of all relevant facts (such as *"slow traffic"* and *"car crash"*)

Location the set of all relevant locations

- **Name** a string attribute assigning a name to the location
- **Geometry** a 2D attribute assigning a geometry to the location

POI the set of all places-of-interest, a specialization of **Location** (such as *"North Shopping*" and "*West Hospital*")

Street the set of all relevant streets, a specialization of **Location** (such as "*Main Street"*)

- **Two-way** a Boolean attribute which is true when the street is two-way

Relationships

occurs relates a fact to a single location, where "*F* **occurs** *X*" indicates that *F* is a fact that occurs in a location *X*, in which case we say that X is the *main location of interest* for *F*

Both a Boolean attribute which is true when *X* is a two-way street and *F* affects *X* in both directions

qualifies relates a street *X* to a location *Y*

How an attribute with one of the following 3 values:

direction indicates that *Y provides a reference direction* for *X* (such as "*Main Street in the direction of the North Shopping*")

restriction indicates that *X is restricted to Y* (such as "*Main Street restricted to the South Borough*")

reference indicates that *Y provides a reference location* for *X* (such as "*Main Street having as a reference the West Hospital*")

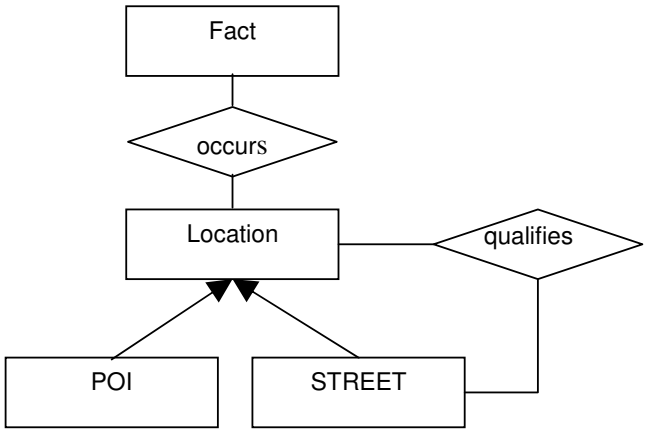

Figure 2: Simplified entity-relationship diagram of the geocoding model.

3.2. A Typical Use of the Model

This section describes the typical spatial operations performed to improve the geocoding of a fact.

Let *F* be a fact that occurs at a location *M*, called the *main location of interest*.

Assume that *M* is restricted by a location *A* and that the geometry of *A* is a polygon. Then, we may use *A* to filter *M* in two different ways: (i) by geocoding the boundaries of *A* and using them to filter *M*; or (ii) by appending the location name of *A* to the main location *M*.

Assume that *M* is a two-way street and that *D* provides a reference direction for *M*. Then, we may call a routing service, passing as parameters *M* as the origin and *D* as the destination, to discover a route *r* that goes from *M* to *D*. Then, we may use *r* to simplify the geometry of *M* to just the affected direction.

Assume that *M* is a street and that *R* provides a reference location for *M*. Then, we may use the geometry of *R* to again simplify the geometry of *M*. For example, if the geometry of *R* is a point (i.e. a building), we may discard those parts of the geometry of *M* that lie outside a circle of a given diameter whose center is the geometry of *R*.

Figure 3 illustrates the result of applying this process to the text example described in Section 2. Section 4.1 further illustrates the process.

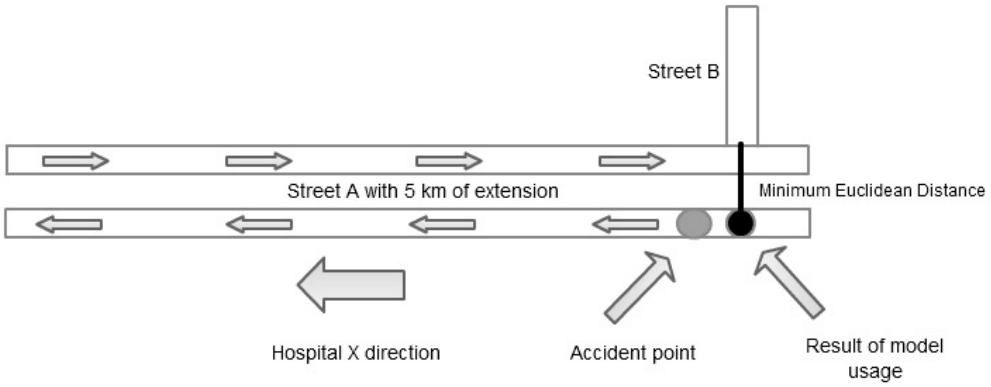

Figure 3. More precise result with the proposed model.

4. Prototype Application

The prototype application implements the process outlined in Section 3 to georeference the locations of traffic-related tweets. This section describes the prototype application and is divided into two parts. The first part describes how tweets are processed, while the second part describes the implementation of the geocoding process.

4.1. Text Data Structuring

Structuring raw text data and extracting relevant information is not a trivial task. The Locus system [Souza *et al*., 2004], an urban spatial finder, has an advanced search feature with a georeferencing objective similar to ours, although with a different implementation. It allows searches with "where" and "what" inputs, similarly to our reference approach. Borges *et al*. (2007) use predefined patterns to extract addresses from Web pages using a set of regular expressions.

In our case, however, using a set of regular expressions, such as an address, a place, a neighborhood or a city to extract locations from raw text would not be very effective. We therefore resorted to Machine Learning techniques dealing with Brazilian Portuguese to assign a structure to traffic-related messages [Albuquerque et al., 2012]. The proposed process to structure raw text data is divided into two parts: (i) identifying relevant entities in the text; (ii) inferring the relationship between these entities to generate a dependency tree. Figures 4 and 5 briefly illustrate these two parts.

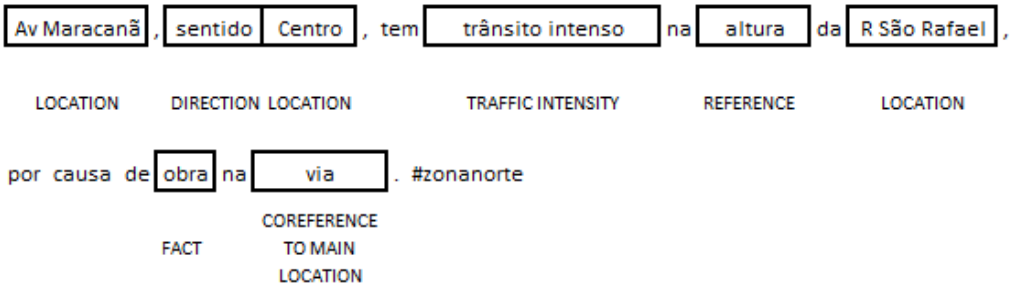

Figure 4. A real traffic-related tweet (in Portuguese) with its entities.

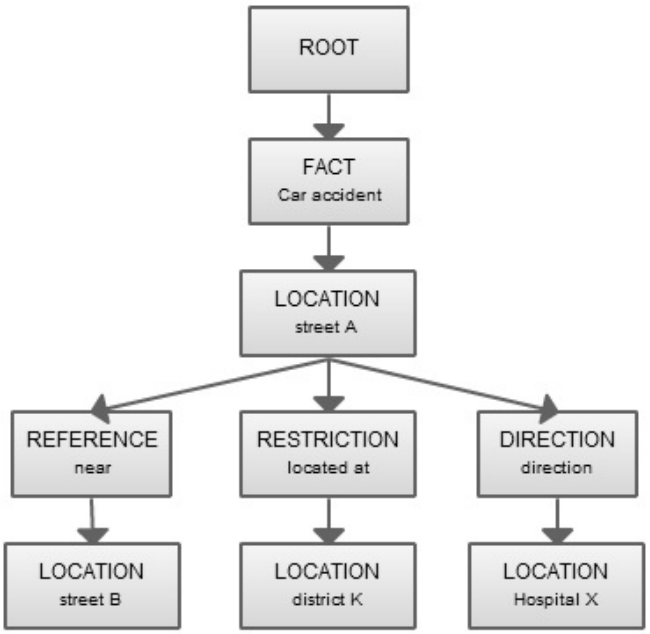

Figure 5. Example of relationship between identified entities.

4.2. Implementation

We implemented the geocoding process outlined in Section 3.2 using services available on the Internet.

We adopted the JTS Topology Suite (Aquino and Davis, 2003), a Java open-source API that implements many 2D geometry functions. Some of these functions and common geometry types are summarized in Bressan and Zhang (2005), which also propose a benchmark for XML processing in GIS applications.

CloudMade and Google provided the geocoding and routing services. CloudMade offers tools and APIs to develop location-based applications, including geocoding and routing services, using the Open StreetMap (OSM) database. An advantage of using OSM is that this service returns the geometry of roads and buildings

(e.g. for a road, it returns a line or multiline and, for a building, it returns either its coordinates or the polygon contour, whichever it is available). The geocoding and routing services provided by Google act as a backup resource: they are used when CloudMade cannot find a valid geometry for the desired geocode location or route. Google's geocoding service does not return geometries when the geocoded object is street-based. This is a problem because it affects the quality of the results.

One common issue in this prototype application is the nature of Twitter text data, which includes abbreviated or hashtag locations (e.g. "Linha Vermelha" is also referred to as "#LinhaVermelha"). To address this issue, we used a synonym dictionary.

Another frequent issue involves classifying certain terms that define a region or a neighborhood. One example is *downtown* (in Portuguese, *centro* or *#centro*), which is often used as a direction but also as a reference. However, since routing services expect addresses or coordinates, we handled this issue by resorting to a particular database of general locations searched before any routing or geocoding operation is invoked.

Consider the following tweets as (real) examples:

(a) *"Acidente entre dois carros no Aterro do Flamengo"*. (*"Accident between two cars at Aterro do Flamengo."*)

(b) *"Acidente envolvendo dois carros no Aterro do Flamengo, sentido #zonasul, na altura da Avenida Oswaldo Cruz."* (*"Accident involving two cars at Aterro do Flamengo, direction #zonasul, near Oswaldo Cruz Avenue."*)

The main location is always associated with a fact. To use this information, we refer to a specific dictionary to identify the type of fact and offer a good visual representation of facts.

Figure 6 shows the results of the analysis of both tweets. Figure 6(A) illustrates the geocoding process without applying the techniques outlined in Section 3.2 (tweet (a)). Figure 6(B) shows the higher precision achieved by applying the techniques of Section 3.2 (tweet (b)), highlighting the correct side of street and the precise location of the accident.

5. Conclusions and Future Work

We described a prototype application that uses traffic-related tweets, in raw text form, to georeference relevant facts over a road network. The prototype takes into account aspects of natural language regarding the description of the location of a fact. The initial results demonstrate that it is indeed possible to retrieve additional data from textual references and use them to improve the georeferencing task. The prototype can be used in applications that monitors moving vehicles in a road network in real-time.

As for future work, we include using a cache strategy to reduce the network traffic overhead caused by the use of the Internet and to avoid exceeding the limits imposed by some Internet services. We also plan to automatically infer fact types, using thesaurus such as WordNet to parse facts from raw text.

(A) **(B)**

Figure 6. Locations extracted from the analysis of tweets.

References

Albuquerque F. da C., Barbosa I., Casanova M. A., Carvalho M. T., Macedo J. A. (2012) "Proactive Monitoring of Moving Objects", Proc. 14th International Conference on Enterprise Information Systems. ICEIS, p. 191-194.

Albuquerque, F. da C., Bacelar F. C., Tapia X. A. C., Carvalho M. T. (2012) "Extrator de Fatos Relacionados ao Tráfego". SBBD - Simpósio Brasileiro de Banco de Dados, p. 169-176.

Borges K. A. V., Laender A. H. F., Medeiros C. B., Davis C. A. (2007) "Discovering geographic locations in web pages using urban addresses". GIR, p. 31-36

Bressan, S., Cuiyu Zhang (2005) "GéOO7: A Benchmark for XML Processing in GIS" Database and Expert Systems Applications. Proc. 16th International Workshop, pp.507-511, doi: 10.1109/DEXA.2005.99

CloudMade, http://cloudmade.com

CloudMade Java Library API, http://developers.cloudmade.com/projects/show/java-lib

Goldberg DW, Wilson JP, Knoblock CA (2007) "From Text to Geographic Coordinates: The Current State of Geocoding". URISA J 2007, 19(1):33-47.

J. Aquino, M. Davis (2003) "JTS Topology Suite Technical Specifications, version 1.4", Vivid Solution, Inc.

JTS Topology Suite, http://www.vividsolutions.com/jts

Souza L. A., Delboni T M., Borges K. A. V., Davis C. A., Laender A. H. F. (2004) "Locus: Um Localizador Espacial Urbano". Proc. GeoInfo, p. 467-478