

# Visualizing Unstructured Text Documents using Trees and Maps

Ian Turton, Alan M. MacEachren  
GeoVISTA Center, Pennsylvania State University,  
University Park, PA, 16802  
{ijt1, maceachren}@psu.edu

## Introduction

Implicit geographic references are common in text documents of many kinds (e.g., field reports, audio transcriptions, human generated route directions). Thus, text sources provide potentially important geographic information that often is under-utilized; in addition to direct references to place names that are commonly extracted and mapped, documents contain rich descriptions of routes, movement, and distributions that (if captured and interpreted) could complement explicit geographic information currently stored in geodata repositories. But, these implicit qualitative geographic references can only be identified and interpreted if put in an appropriate context. While humans are quite good at determining context for interpretation of text, there are orders of magnitude too few human analysts to process all text documents relevant to real-world applications in which the web and other sources provide thousands or hundreds of thousands of potentially relevant documents. Thus, the work outlined here is part of a larger project focused on developing strategies and implementing methods for improving geographic information retrieval and interpretation from text through methods for contextualizing the interpretation process.

The focus of this paper is on developing geovisual analytics tools that support both analysts tasked with identifying documents that describe movements of interest and researchers working to develop understand spatial language leading to computational methods for extracting geographic statements about movement from text. The computational methods focus on combining natural language processing methods for interpreting text with geographic information retrieval strategies for contextualizing that interpretation and mapping services for representing the result. Specifically, to aid analysts and researchers with the tasks involved in identifying, interpreting, and mapping statements about movement in text, we have developed a new hybrid tool that allows the visualization of unstructured text documents as TermTrees and links these text representations directly to maps through the geographic elements that have been identified or reduced to a small number of choices using heuristic methods.

## Methods and Preliminary Results

The initial system for addressing the challenges outlined above is composed of a collection of separate programs for development purposes that will ultimately be integrated into an analytical workbench for processing and geographically contextualizing text documents. The overall system is currently designed to work with unstructured and semi-structured text documents containing route descriptions. An example of a representative semi-structured document is shown in Figure 1. In this case, the file represents the output from a separate program in the document processing chain which converts raw web pages into the XML file shown.

This paper focuses specifically on the component of the system that is designed to linguistically and geographically contextualize spatial terms used in route directions (or other text passages containing spatial references). This part of the system includes two dynamically-linked components: a **TermTree** to explore use of terms and phrases in linguistic context and a **RouteSketcher** to provide the geographic context needed to interpret ambiguous spatial references. Each is described below briefly, along with an example of their interaction.

```

<Routes>
  <Route>
    .....
  </Route>
  <Route>
    <Destination>
      Directions to Reproductive Science Institute &
      Valley Forge Surgical Center 945 Chesterbrook
      Blvd. Chesterbrook , Pennsylvania 19087
    </Destination>
    <Origin>
      From Philadelphia , Schuylkill Expressway West (I-76)
    </Origin>
    <RouteDesc>
      Take I-76 West (toward Valley Forge) Take Exit
      #328A to 202 South Take the Chesterbrook Exit
      to stop light Turn right onto Chesterbrook
      Blvd. & merge into the left lane At first
      light , Duportail Road, turn left Make the first
      left into parking lot Building is on the left.
    </RouteDesc>
  </Route>
  <Route>
    .....
  </Route>
  <Route>
    .....
  </Route>
</Routes>

```

Figure 1: Part of the Route File

The TermTree is currently able to visualize a document (or set of documents) by allowing the user to select a word and then see a pair of trees representing all phrases that end and start with the selected word (see figure 2). The user can sort these trees by frequency or alphabetical order as required. This visualization allows users to select common phrases that contain a word of interest (as illustrated for the term “left”). The system also applies preprocessing methods to reduce the variations in the phrases presented, as the current system is geared to working with route directions the current preprocessing reduces highway and interstate names to a generic route tag and numbers to a number tag so that the phrases “go 5 miles to I-80” and “go 8 miles to Interstate 95” are both shown as “go NUMBER miles to ROUTE”. This allows the grouping of semantically similar phrases that differ in the exact number of miles for example or that refer to different routes.

In the near future we intend to extend the preprocessing steps to include the improved street name recognition algorithms developed as part of the geocoding work described below. The system will also be improved by the addition of the ability for a user to enter a phrase rather than a single word to use as the root of the visualization. This will allow the exploration of more interesting groups of phrases in groups of documents. A further development will be to allow the use of wild cards in the selection term, so that the user could select “turn \* at” to match “turn left at” and “turn right at”.

The RouteSketcher part of the system parses the same XML file containing the semi-structured route description into a series of geographic way marks that can be displayed on a map for the user to inspect. The objectives for this tool are:

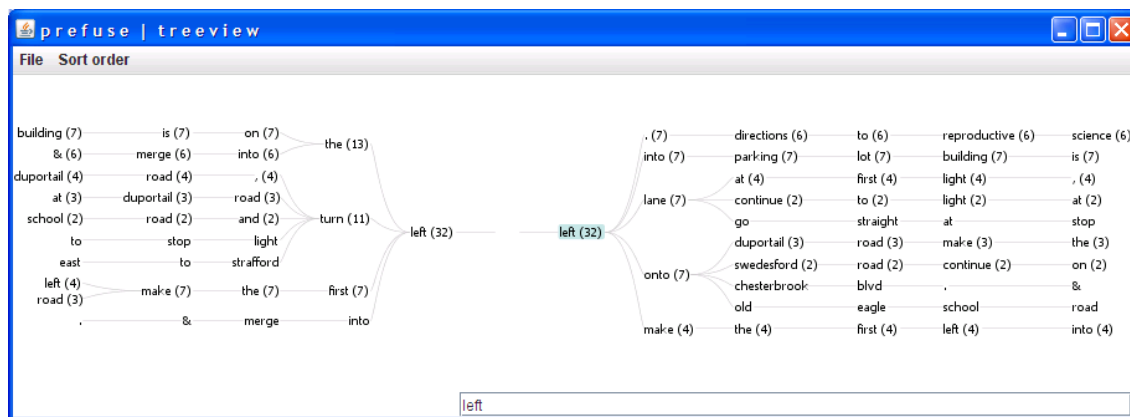


Figure 2: The TermTree tool showing the route description

- (a) identify all definite or possible references to geographic entities (e.g., roads, intersections, regions, buildings, parks, etc.),
- (b) using relevant databases geocode the entities,
- (c) display the geocoded entities on the map display with indications of certainty,
- (d) support user interrogation of the result through dynamic links to the TermTree and to the document, and
- (e) support user feedback to correct and/or refine system decisions.

The system first reads in the XML file and parses the contents of the file to extract origin, destination and route instructions. Each of these strings is then tokenized to words, punctuation and white-space; the processor then passes through the token stream extracting geocodable tokens such as zip codes and telephone numbers that can be recognized by using a simple regular pattern matching expression. Each text token that matches is changed to a more specific token type to reflect the identification. Then each token is examined to see if it matches the program's list of possible street name suffixes (generated from the US Census TIGER data set) e.g. Road, Blvd, St, etc. Once a possible suffix has been identified the program works backwards through the token stream adding tokens that represent white-space or that start with a capital letter (indicating a proper noun), when this search is complete the program combines the tokens found into a new token of type street and replaces the other tokens in the stream with this new one. Once street names have been identified the system looks up other capitalized words (or groups of words) to see if they are towns or cities using the GeoNames system.<sup>1</sup> Finally a pass is made through the token stream to extract addresses which are defined as all the tokens between a number followed by a street token until a zip code token, when found these groups of tokens are combined to form an address token. Once the text has been tokenized fully each special token is looked up using a specific web service or local database to acquire a set of geographic coordinates for it. Addresses are geocoded,<sup>2</sup> populated places are looked up<sup>3</sup> and roads are resolved using a local database of the Open Street Map data. However as the entities under investigation are streets which, while there are sometimes comprehensive databases available, are rarely unique. For example there are nearly 13,000 roads called Main St in the continental US and this number rises to 28,500 when geographic variants like North Main St are considered. This repetition of street names increases the complexity of the geocoding problem over the related problem of place name disambiguation where even common place names like Springfield only occur 910 times. Fortunately the task is eased in that documents that mention street names often mention groups of streets that are near each other or specifically mention a town name with the street name. So by applying spatial

<sup>1</sup><http://www.geonames.org>

<sup>2</sup><http://geocoder.us>, currently only US addresses are handled

<sup>3</sup><http://www.geonames.org>

