09-03

On-demand Geo-referenced TerraFly Data Miner

Naphtali Rishe, Maxim Chekmasov, Marina Chekmasova, Scott Graham, Ian De Felipe
High Performance Database Research Center
Florida International University
11200 SW 8 ST, ECS-243
Miami, FL 33199 USA
+1 305 348 1706
{rishen, maximc, marinac, grahams, idefel01}@cs.fiu.edu

1. ABSTRACT

We present a comprehensive Internet data extraction tool adopted for the TerraFly Geographic Information System (GIS). TerraFly is a web-enabled system that allows users to virtually fly over remotely sensed data, including satellite imagery and aerial photography, using a standard Internet browser. The data extraction tool presented here is designed to augment the user's virtual flight experience with extensive data relevant to any given geographical point along the virtual flight path. The data presented to the user is retrieved from several server-side databases and is collected from the Internet data providers using our patented data extraction technology. Some data elements are presented to the user as overlays, some in popup windows, and some via hyper-linking to third-party web sites.

Categories and Subject Descriptors

D.3.3 [**Information Storage and Retrieval**]: On line Information Services – Web-based services

General Terms

Algorithms, Design, Human Factors, Management, Performance

Keywords

Geographic information system; remotely sensed data; Internet data extraction.

2. INTRODUCTION

The availability and use of remotely sensed data has steadily increased since the first commercial imaging satellite was launched several years ago. To meet an increasing demand for this data, the TerraFly system has been developed. TerraFly is an interactive fly-over vehicle designed to aid in the visualization of remotely sensed and spatial data via the Internet. The system has a broad range of spatial data manipulation capabilities. TerraFly's features include:

- Smooth flight over spatial data: streaming incremental tiles to a Java applet allows users to virtually fly over available data.
- Synchronized flight: multiple frames allow users to fly in sync over different data sets.
- Compass Control: supports 360 degree of flight direction and speed. It is also used to control data refresh rate.
- Zoom In/Zoom out: allows the user to view spatial data at varying resolutions.
- Image Processing Filters: allow users to customize the display of data. Filters enhance the appearance of images.

We refer to [5] for further information about the TerraFly system.

While TerraFly gives users a unique visualization experience by flight simulation, a large number of its applications require additional data to be made available to the user alongside the imagery. This brings to life the concept of data mining based on geographical location. More precisely, each geographical location is considered as a point identified by coordinates (say, by Easting, Northing, and Zone in a Universal Transverse Mercator (UTM) grid). By clicking a geographical point on the imagery, the user initiates a data mining process against numerous data sources to get information on the point and the area(s) the point belongs to. The data to be collected and presented to the user is diverse: demographic, social, economic, and environmental among others. The granularity of data is also variable. For example, the point clicked within the continental United States lies in some block group, census tract, zip code, county, and state. Each of these census designated areas has its own demographic dataset with specific format of data presentation.

3. DATASETS AVAILABLE TO THE USER

The largest and most comprehensive dataset currently used in the data mining module is collected from the US Census Bureau, see [6]. The US Census dataset contains both textual and cartographic information and is based on the census conducted in the United States in the year 2000. The user gets demographic statistics for the following areas which contain the geographical point clicked: Block Group, Census Tract, Zip code, City, Census County Division (CCD), Congressional District, County, Metropolitan Area, and State. A quick look report gives the user area population, square mileage, number of housing units, and percentage of territory covered by water.

A mouse click on any of the displayed areas delivers a comprehensive demographic report to the user. This report provides detailed figures on population, including race, age, sex; on households including size, type, presence of children, non-relatives; on families including type, size, race; imputations of sex, race, age by housing units including urban/rural, occupancy/vacancy status; and other information. The data mining tool supplements the demographic report with an Expanded Map

IUI'03, JANUARY 12–15, 2003, MIAMI, FLORIDA, USA.
ACM 1-58113-586-6/03/0001.

Album, containing the following maps encompassing the point of interest:

- Far ZoomOut map;
- Regional map;
- Street Detail map;
- Median Family Income ZoomOut map;
- Population Density ZoomOut Map.

The Geographic Names Information System (GNIS) database contains information about almost 2 million physical and cultural geographic features in the United States, see [3]. It is used as another data source for TerraFly's on-click data mining tool. When the user clicks on the imagery to identify the geographical point, she gets a list of the five nearest objects retrieved from the GNIS database. The distance in feet or miles and a direction arrow indicating the proximity of the object to the point clicked is calculated by the data mining module and is displayed along with the name of the object. Parks, rivers, bridges, distinguished buildings and monuments are some examples of GNIS objects. The names of the objects are also clickable for further data mining: a mouse click initiates a search for additional information about the object using the Google search engine, see [4]. However, to reduce a number of returned web pages related to the object and to force the search output to be more geographically specific, the data mining module calculates the zip code of the object area and passes it to Google as an additional search parameter. A "More" button is presented to allow the user to get 100 nearest GNIS objects from the point clicked.

Proven to be important for travel applications, a Hotel dataset is also available for data retrieval. Presentation of hotel data to the user follows the same rules as GNIS objects: distance and direction arrow from the geographical point clicked and a "More" button for a larger set of records. However, unlike the GNIS set, the hotel dataset provides detailed information. Besides the hotel's name, a description of the hotel facilities, policies, price range, address, and indoor/outdoor photos are displayed for the user to easily compare nearby hotels and to make a decision on hotel preference. Hotel reservation is one button click away with the user invited to specify dates to check availability and further process of booking.

A similar approach is used to display real estate properties for sale in the vicinity of the point of interest. General and real estate specific data is mined for the user: property overview including price, square footage, number of baths, year built; property description including photo, address, community information, property features; contact information, including real estate agency name, address and phone numbers. A uniform resource locator (URL) is computed to lead the user to the particular property on the data provider's web site, where the original data is not ameliorated and has a different look-and-feel.

School and crime data may be viewed by the user in addition to real estate properties. This data characterizes neighborhoods to a large extent. General information on each school includes name, address, contact phones, type (public or private), grade levels, total number of students and number of students per teacher. Detailed school information contains enrollment by race/ethnicity and by grade, various statistical figures, school identifications, and programs. Crime data is related to the zip code of the clicked geographical point. The major characteristic is crime rating, which is a weighted average for the entire zip code.

TerraFly's on-click data mining module provides access to several datasets related to natural and environmental resources. A simple feature displays current time and weather conditions in the area of the geographical point clicked. Additional historical weather and basic astronomical data for the region is available via the URL. The Federal Emergency Management Agency supplies the data mining module with hazard information, most notably hazard maps, see [1]. Floods and earthquakes data are of particular importance for any region. The fire detection program of the National Oceanic and Atmospheric Administration is a data source for urban and forest fires, [2]. After processing of the fire dataset, TerraFly's module displays date/time, coordinates, direction, temperature, land cover and distance to the point clicked by the user assembled in one table.

To easily correlate the above mentioned datasets with the imagery, the data mining tool displays a static aerial photography image of the area encompassing the point clicked. The user is able to instantly change the resolution of the aerial photography by zooming in and out in the range of 1-64 meters per pixel. The image contains several layers of data:

- icon of the geographical point clicked,
- street names,
- highways,
- hotel icons and names.

The user may "travel" in any of eight compass directions to adjust or further explore imagery beyond the initial static image.

4. DESIGN OF THE DATA MINING MODULE

TerraFly's data mining module has been designed to easily add and remove data sources for the application. Hypertext transfer protocol (HTTP) is adopted as the internal interface between the server side and the data providers. All data requests are sent via HTTP using URLs. Most data is returned to the client side in hypertext markup language (HTML) or plain text formats.

Three types of data providers may be identified for the data mining module, namely:

- Third party service. The data mining module calculates the parameters of the URL to the third party web site. Thus, the URL with some particular data request is sent to the external data provider.
- TerraFly's internal data provider. Internal providers are usually implemented on a separate physical server with web server software installed and running. The Common Gateway Interface (CGI) programs reside on the server to receive data requests, query the database and send query results back to the client side. In this case the database is stored in-house; it usually resides on a separate database server to provide easier system maintenance.
- TerraFly's data service. The data service receives the requests, compiles and sends web agents to collect data from the Internet in real time. We refer to [7] for the description of this technology.

The server side of the data mining module is yet another CGI program. When the user clicks a geographical point in TerraFly,

TerraFly calls the CGI program with the coordinates of the point and other parameters. The server side program calls providers for the initial data and generates the HTML page to be presented to the user. Based on the user's preferences and data mining activity, the program calls the corresponding data providers for additional datasets.

One internal service is not directly seen by the user, but is heavily used by the data mining server side program. This is the spatial service, which is responsible for conversion of geographical coordinates into place names and vice versa. This service performs the following functions:

- Depending on the request, given geographical coordinates are transformed to street address, zip code, name of the nearest city, and so forth.
- Given street address it returns geographical coordinates. Given name of geographical area (like zip code, county, state) returns official coordinates of the center of the area.
- Given rectangular geographical area it returns zip code(s), district/city/county/state name(s), associated with this area.

The importance of the spatial service is that it allows the integration of textual data with remotely sensed imagery.

TerraFly displays remotely sensed imagery and is designed to work in terms of geographical coordinates. Most of the sources return data in terms of addresses and place name with no geographical coordinates. The spatial service supports association between names and coordinates thus allowing TerraFly together with its data mining module to provide the user with a unique application.

Spatial service is implemented using algorithms described in [8]. Being designed for specific conversion needs, the service is exceptionally fast, handling up to several thousand requests per second.

5. CONCLUSION

We have discussed an on-click data mining module for TerraFly GIS. Data mining is organized for numerous data sources. The design of the system allows the easy addition of new datasets. Additional data provided by the module significantly enriches the user's experience with the TerraFly GIS. This data mining module may also serve as a prototype to a series of specialized TerraFly applications, including real estate, travel and emergency response preparedness.

6. ACKNOWLEDGEMENTS

This research was supported in part by NASA (under grants NAG5-9478, NAGW-4080, NAG5-5095, NAS5-97222, and NAG5-6830), NSF (EIA-0220562, CDA-9711582, IRI-9409661, HRD-9707076, and ANI-9876409), the Department of Interior, and ONR.

7. REFERENCES

[1] Federal Emergency Management Agency. Available at http://www.fema.gov/.

[2] Fire detection program, National Oceanic and Atmospheric Administration. Available at http://nhis7.wwb.noaa.gov/website/SSDFire/viewer.htm

[3] The Geographic Names Information System, United States Geological Survey. Available at http://gnis.usgs.gov/.

[4] Google search engine. Available at http://www.google.com.

[5] TerraFly: A Web-Enabled Application for Visualization and Manipulation of Remotely Sensed Data, white paper, http://terrafly.fiu.edu/tf-whitepaper.pdf

[6] US Census Bureau, United States Department of Commerce. Available at http://www.census.gov/.

[7] Rishe, N. Data extractor, U.S. Patent 6339773. Issued January 15, 2002.

[8] Shaposhnikov, A., and Rishe, N. S-tree - a High Performance Multidimensional Spatial Index, HPDRC internal report, May 2002. Available via contact at http://hpdrc.cs.fiu.edu/.