GEDMWA: Geospatial Exploratory Data Mining Web Agent

Edward Pultar
University of California, Santa Barbara
5807 Ellison Hall, Dept. of Geography
Santa Barbara, CA 93106-4060
[001] + 805.893.4519
pultar@geog.ucsb.edu

Martin Raubal
University of California, Santa Barbara
5713 Ellison Hall, Dept. of Geography
Santa Barbara, CA 93106-4060
[001] + 805.893.4839
raubal@geog.ucsb.edu

Michael F. Goodchild
University of California, Santa Barbara
5707 Ellison Hall, Dept. of Geography
Santa Barbara, CA 93106-4060
[001] + 805.893.8049
good@geog.ucsb.edu

ABSTRACT

An abundance of geospatial information is flourishing on the Internet but mining and disseminating these data is a daunting task. With anything published on the web available to the public it has become a grand repository of volunteered geographic information (VGI). Internet users often provide location information for videos, pictures, travel destinations, or other events. All of these data can be gathered by a web crawling *geospatial agent* that later performs geospatial data mining. The discovered geoinformation can be stored, analyzed, queried, and visualized as the agent creates a data repository of what it discovered. This paper presents the design and prototypical implementation of the GEDMWA (Geospatial Exploratory Data Mining Web Agent). It reads webpage data and follows links to acquire knowledge in order to add value to geoinformation usable in a GIS. The agent creates a database of webpage text, mines it for location information, and then converts it to proper geospatial data format. The data is quickly visualized and analyzed after GEDMWA converts it into proper GIS and virtual globe formats. This provides diverse user communities a tool that utilizes a variety of distributed sources to discover additional knowledge about their fields of interest.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Geospatial Agents – *retrieval models, search process, selection process.*

General Terms

Management, Measurement, Design, Experimentation.

Keywords

Geospatial agent, VGI, geographic data mining.

1. INTRODUCTION

The amount of geospatial information available through the Internet is ever-increasing but discovering the data and converting it to a useful format is a challenge. This is because of the unique combination of techniques and the necessary communications between the components. An increasing number of users are volunteering geographic information via means such as Google Earth and Wikimapia[1] [1]. However, as anything published on the web is available to the public it is a repository of VGI. Internet users often provide location information, e.g., latitude / longitude coordinates for a video, a picture, travel destination, or other data. A web crawling *geospatial agent* can acquire this information and perform geospatial data mining. Using open source software solutions such as those supported by the OGC[2] the geoinformation can be stored, analyzed, queried, and visualized as the agent creates a data repository of what it discovers.

This paper focuses on the development, prototypical implementation, and testing of such geospatial agent: GEDMWA (Geospatial Exploratory Data Mining Web Agent) can read webpage data, follow links, and acquire spatial information to be used in a GIS. Data exploration using this agent can discover previously known properties of a phenomenon and add value to geographic locations through explicit location-based semantic data mining of the Web. The resulting information can be quickly visualized and analyzed after GEDMWA converts it into proper GIS and virtual globe [2] formats such as the Keyhole Markup Language (KML). A wide variety of user communities can then use a diversity of international sources to discover additional knowledge about the field of interest.

The proposed geospatial agent functions using VGI put on the Web by the general public. This creates a large dataset with many potential data providers across the globe. It also brings into question who places geographic data on the web and what is the associated quality of data [3]. This agent can be used to learn what types of geodata are posted on the Internet, i.e., latitude & longitude, KML, state plane projections, and UTM among others.

The goals of this research include:

1. Create a geospatial agent to locate VGI on the Web.
2. Produce valid geospatial data (KML, shapefiles, others) from these discovered data using open source software solutions.

2. USE CASES

The proposed geospatial agent has utility for various user communities around the globe. Take for example a group of mountain climbers. A blunt way of using geospatial data to find new areas to climb would be examining topographic maps and looking for suitable changes in elevation. However, this may not always yield desirable results as blank mountain faces with no holds can be impossible to climb. Members of this community

ACM GIS '08 , November 5-7, 2008. Irvine, CA, USA (c) 2008 ACM
ISBN 978-1-60558-323-5/08/11...$5.00

[1] http://earth.google.com; http://www.wikimapia.org

[2] http://www.opengeospatial.org

post information on the web with geographic coordinates and descriptions of climbing routes. With the use of a GPS receiver the general public can collect location data in the form of latitude and longitude among others. Gathering this VGI using Geographic Information Retrieval (GIR) is possible using the geospatial agent developed in this research.

The GEDMWA can be initially directed to pages on the web that contain the phrase "rock climbing". A user can then specify a number of pages to consider in the creation of this spatial database in order to limit the search. The Java^{TM}-based agent searches for patterns of coordinates in the pages and creates the associated geospatial data files. All of these georeferences can be compiled into one source and easily imported into a virtual globe environment or open source GIS, such as Quantum GIS or uDig[3].

Assembling these data together adds a utility value to the discovered spatial locations for each member of the community. Additional use cases are found in a variety of communities such as those concerned with geysers, historical locations, landmarks, animal watching [4]. GEDMWA is able to locate and describe what exists in the physical world by using the virtual environment of the Internet. This aids in the decision-making capabilities of users as they gain a better knowledge about locations of interest.

3. RELATED WORK

An agent can be regarded as anything that perceives its environment through sensors and acts upon that environment through effectors [5]. More specifically, agents are considered computer systems that are situated in some environment and can act autonomously [6]. Agents have been mainly dealt with in Artificial Intelligence but have recently also gained popularity in other fields such as geography [7] [8]. An extensive review of the current state of these geospatial agents is provided by Sengupta *et al.* (2008) [9].

VGI harnesses the power of having geoinformation authors across the globe. The Internet aids in sharing this information publicly. The growth in VGI has quickly increased in recent years and continues into the future as an element of Web 2.0 [10]. The system presented in this paper utilizes VGI as a key component for creating geodata that are visualized in a GIS or virtual globe environment. The web as a source of VGI has been realized for modeling vague places [11] in the UK as well. It is an excellent example of the potential the web has for GIScience as it generates boundaries and surfaces of regions based on a place name.

Turning web data into usable information is a task that can utilize methods such as simple data retrieval or knowledge discovery. Data mining [12] and its tools are critical parts of the proposed geospatial agent. Geographic knowledge discovering and spatial data mining are tailored for extracting information and finding interesting patterns in databases with a spatial component. Miller [13] describes this in further detail along with other key techniques. Specifically searching for information through data mining the web is explained in detail by Chang *et al.* [14].

4. METHODS

GEDMWA utilizes a number of tools and steps to complete its tasks. This includes VGI, the Internet, data mining and creation, and GIS. The agent's overall flow of data, methods, and tools is visualized in Figure 1 and described in the rest of this section.

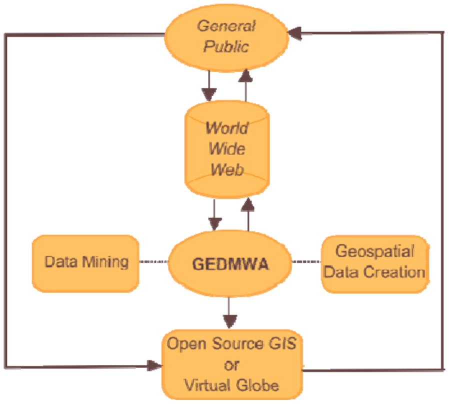


Figure 1. GEDMWA Data Inputs and Outputs

Initially GEDMWA crawls the web from a starting URL searching for public geospatial data. This could be a topic search and the page returned by a Google query. It may be any valid URL and, in the interest of discovering geospatial data, one that is likely to have coordinates. The agent downloads the content of the page using Java™ methods for URLs

```
URL searchURL = new URL("http://wikipedia.org");
    URLConnection urlConn = url.openConnection();
      InputStream urlStream = url.openStream();
```

After the connection is opened a read method is used to acquire the entire page text. Valid links are saved for further investigation and later followed recursively (abiding by the Robots Exclusion Standard[4]) until they reach the user-specified limit. A basic text database is used for testing the prototypical implementation where each URL's text is stored for later geographic data mining purposes. Future versions can support SQL queries utilizing more rigorous, open source database software such as HSQLDB[5].

Once the agent has collected the desired amount of pages it can begin the dissection and data mining process. The key part of this process involves the search for geographic coordinates. The Java^{TM} programming language provides the powerful pattern-matching tool of *regular expressions*, which can be utilized to locate geographic coordinates within each page. *Regular expressions* allow for automatic retrieval of text strings containing sequences such as "dd.ddddddd" where each d corresponds to a digit: 0 thru 9. E.g., the following code would find latitude decimal degrees coordinates in a form similar to "N 46.853441":

```
[NS]\\s\\d{1,2}\\.\\d{3,10}
```

This searches text for an initial N or S (North or South) character, then a space before 1 or 2 digits followed by a dot (.) ending with 3 to 10 digits of precision. Similar expressions can be used to find longitude coordinates as well as coordinates in other formats. As

[3] http://www.qgis.org; http://udig.refractions.net

[4] http://www.robotstxt.org

[5] http://hsqldb.org

another example, coordinates in the form of degrees, minutes, and seconds (e.g., N38° 35' 47") can be discovered by using the following code:

```
[NS]\\d\\d°\\s+\\d\\d'\\s+\\d\\d\"
```

This searches for either an initial N or S character followed by any two digits, the degree symbol, any amount of whitespace (space, tab, newline, other), any two digits, a single quote, any whitespace, any two digits, then finishing with a double quotation mark. The utility attained by harnessing the power of *regular expressions* quickly becomes apparent as many different forms of geoinformation can be discovered with little code.

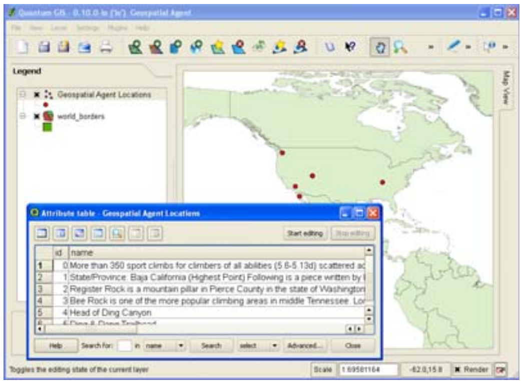
Once regular expression code is written in proper syntax it provides an efficient way to search the large text databases the agent has retrieved from the web. Text surrounding any discovered coordinates can be stored and attached to the location as an attribute. The amount of text stored in the attribute is defined by the user with various combinations tested for the initial implementation of the agent. E.g., one may specify to use the 500 characters preceding a discovered coordinate or the 500 characters following the coordinate. Storing this information provides details on the context of the geospatial data including any available descriptions and metadata. The data can later be transformed into a standard geospatial format such as KML or a shapefile.

The transition from the discovered text-based geographic coordinates to proper geospatial data is aided by the use of the open source GIS toolkit GeoTools[6]. Written in Java™, GeoTools provides helpful methods for converting coordinates into geospatial data formats such as those officially accepted by the OGC. Code for KML output was custom-made for this research (although using GDAL[7] can provide even more options for geodata formats) and can be imported into a virtual globe as discussed in Section 5. The shapefiles created by this tool can be opened by an open source GIS such as QGIS or uDig.

5. APPLICATION

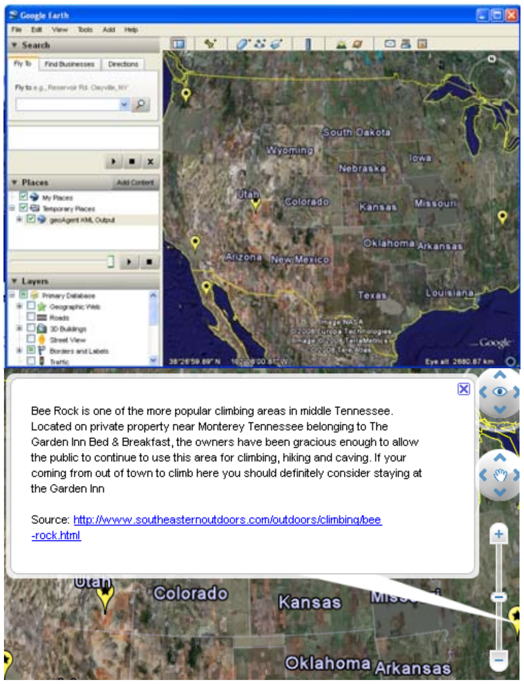
The application was designed using the Eclipse[8] IDE version 3.3.1. Initial results were found using the agent with various starting web locations. Internet-based repositories of geospatial information were found for the various user communities (see Section 2). The user community of climbers is explored in more detail using results from the Google search engine.

The agent created with this paper retrieved resources found by searching for "latitude longitude climbing", "climb crag latitude longitude", and other phrases. The mass of data collected was then investigated for locations using the *regular expressions* methods described in the previous section. Coordinates were discovered in both decimal degrees format (e.g., 38.596389, -110.843611) and degrees, minutes, seconds format (e.g., N 36° 07'19", W 85° 17'07"). Whenever a location was found in a page the coordinates and text near the coordinates (e.g., 300 characters) were saved. This value is specified by the user allowing customization as desired. Once the coordinates and attributes were discovered the agent began the geodata creation process. In this example all of the discovered data were then used to create new GIS data (Figure 2) and virtual globe data (Figure 3).


Quantum GIS - 0.10.0-Io ('Io') Geospatial Agent
File View Layer Settings Plugins Help
Legend
Geospatial Agent Locations
world_borders
Map View
Attribute table - Geospatial Agent Locations
Start editing
Stop editing
Help
Search for:
in
name
Search
select
Advanced...
Close
Toggles the editing state of the current layer
Scale
1:69581164
-62.0,15.8
Render

	id	name
1	0	More than 350 sport climbs for climbers of all abilities (5.6-5.13d) scattered ac
2	1	State/Province: Baja California (Highest Point) Following is a piece written by I
3	2	Register Rock is a mountain pillar in Pierce County in the state of Washington
4	3	Bee Rock is one of the more popular climbing areas in middle Tennessee. Lo
5	4	Head of Ding Canyon

Figure 2. Agent's shapefile results in Quantum GIS.



Bee Rock is one of the more popular climbing areas in middle Tennessee. Located on private property near Monterey Tennessee belonging to The Garden Inn Bed & Breakfast, the owners have been gracious enough to allow the public to continue to use this area for climbing, hiking and caving. If your coming from out of town to climb here you should definitely consider staying at the Garden Inn

Source: http://www.southeasternoutdoors.com/outdoors/climbing/bee-rock.html

Figure 3. Agent's KML results in Google Earth

In the above figures a selection of resources discovered in the Americas is visualized. Additional information about each location is stored in an attribute table for the GIS data and in a description tag for the KML data. In creating the KML data a new Placemark Feature with associated Geometry is made for each location the agent discovers. Once the data files are created the user is able to perform any further operations desired using their tools of choice.

6. DISCUSSION

The majority of the computation time for this agent depends on the speed of data retrieval from the web. Given that this agent can follow links from one country or continent to another the amount of time an execution takes is highly variable. Times varied from a few seconds to many minutes in test runs with the prototype. The

[6] http://www.refractions.net/products/geotools

[7] http://www.gdal.org

[8] http://www.eclipse.org

length of time it takes to retrieve data depends on the bandwidth of the system running the agent. The rest of the Java™ code utilizes *regular expressions* and basic input and output (I/O) techniques inheriting the efficiency of these methods as they are implemented in the language.

Quantification of data quality is a valid topic to be addressed in this context. This relates to quantifying the amount of value added to a discovered piece of geoinformation and assessing the data source. Identification of IP addresses can be used to determine general locations of persons providing VGI. These sources (e.g., Internet cafes, libraries) lend a starting point for establishing levels of data quality that pertain to particular user groups.

Determining the validity of discovered information for particular users is a task that can be approached from multiple directions. One initial method involves using cross page correlation where user communities specify a quantity that is a critical threshold for hinting at valid geospatial information. E.g., the user can specify that any discovered location and keyword pair be found at least 5 times before being stored in the database. Given issues of scale and resolution a buffer range can be specified as well, e.g., the 5 or more data entities must be within a circle of radius 100 meters to be considered appropriate for the user's dataset. These values vary between user groups and may be tweaked to a level that is most useful by the particular people searching for new knowledge.

With this agent implemented one can begin to see how such a system can add value by synthesizing the collected data coordinates with useful information. An essential characteristic of deciding how much value the agent can add is whether or not the integrated data affects a decision [15]. Other variables to consider in quantifying the amount of value added to a location is the quantity of people in a specific user group in addition to the size and quantity of the database created for new locations. Currently for each location the coordinates, source page, and description information are gathered but adding additional attributes such as the amount of links to other pages can be used as a component of measuring and quantifying added value. Allowing the variability in text collected from a page gives a tool for analyzing how much data is needed to create useful information.

7. CONCLUSIONS AND FUTURE WORK

This paper provides an implemented tool for searching the web for location data, amassing these data and mining them for coordinates, then creating new pieces of geodata. The amount of information gained varies between user groups. However, this agent does not retract value but rather adds value to any location data it discovers and creates a central repository available for further analysis. This research is also usable for exploration of the various types and quantities of geoinformation that users make available on the web. Finding common geospatial data formats published on the web will make this tool even more useful to additional user communities gaining utility from this research.

Adding a temporal extension to this agent would deliver both geospatial and time information pertaining to how up-to-date the created data is. This adds another factor for the relevancy and value of the added information. Utilizing HTTP headers retrieved via an HTTP request of a URL provides a value "Last-Modified" that tells how recently a page has been updated. The agent can then store this information with the geospatial data making it available for use in the analysis.

This work continues into the future with support for different types of geoinformation available on the web, such as Universal Transverse Mercator (UTM) and state plane coordinates. The agent developed here demonstrates the abilities attainable by combining geospatial agents with VGI, web crawling, and geospatial data mining and creation.

8. REFERENCES

[1] Goodchild, M. 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211-221.

[2] Grossner, K., Clarke, K. and Goodchild, M. 2008. Defining a Digital Earth System. Transactions in GIS 12, 145-160.

[3] Shi, W., Fisher, P. and Goodchild, M. (Eds.) 2002. Spatial Data Quality. Taylor & Francis, New York.

[4] Larson, K. and Craig, D. 2006. Digiscoping vouchers for diet studies in bill load holding birds. Waterbirds 11, 110-112.

[5] Russell, S. and Norvig, P. 2002. Artificial Intelligence: A Modern Approach. Prentice Hall.

[6] Wooldridge, M. 1999. Intelligent Agents. In Multiagent Systems - A Modern Approach to Distributed Artificial Intelligence, G. Weiss Ed. MIT Press, Cambridge, MA, 27-77.

[7] Batty, M., Desyllas, J. and Duxbury, E. 2003. The discrete dynamics of small-scale spatial events: agent-based models of mobility in carnivals and street parades. International Journal of Geographical Information Science 17, 673-697.

[8] Benenson, I. and Torrens, P. 2004. Geosimulation - Automata-based modeling of urban phenomena. Wiley, Chicester, England.

[9] Sengupta, R. and Sieber, R. 2007. Geospatial Agents, Agents Everywhere... Transactions in GIS 11, 483-506.

[10] Goodchild, M. 2007. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. International Journal of Spatial Data Infrastructures Research 2, 24-32.

[11] Jones, C., Purves, R., Clough, P. and Joho, H. 2008. Modelling Vague Places with Knowledge from the Web. International Journal of Geographical Information Science, in press.

[12] Han, J. and Kamber, M. 2006. Data Mining: Concepts and Techniques, Second Edition. Morgan Kaufmann, San Francisco.

[13] Miller, H. 2007. Geographic data mining and knowledge discovery. In Handbook of Geographic Information Science, J. Wilson and A. Fotheringham Eds. Blackwell.

[14] Chang, G., Healey, M., McHugh, J. and Wang, J. 2001. Mining the World Wide Web: An Information Search Approach. Kluwer, Norwell, MA.

[15] Frank, A. 2003. Pragmatic Information Content - How to Measure the Information in a Route Description. In Perspectives on Geographic Information Science, M. Goodchild, M. Duckham and M. Worboys Eds. Taylor & Francis, London.