MetaData Retrieval: A Software Prototype for the Annotation of Maps with Social Metadata

Rosa Meo[1], Elena Roglia[1,2], and Enrico Ponassi[1]

[1] Dep. Computer Science, University of Torino, Italy
[2] ITHACA, Torino

Abstract. MetaData Retrieval (MDR) is a software module for the enrichment of geo-referenced maps with metadata. Metadata are annotations on spatial locations that are taken from the Volunteered Graphical Information projects like OpenStreetMap and GeoNames.

The MDR user acts with a user-friendly GUI, a Query By Example in which the user specifies in a multi-dimensional data model the spatial objects for which new information are searched for. The request is translated into SQL queries for the database and in web service requests for OpenStreetMap and GeoNames. Downloaded annotations are checked and compared with the history for duplicate elimination. Annotations are presented to the user in the context of an interactive, geo-referenced map and in a hierarchical, ontological structure, that is a facility for indexing and browsing. On demand, an annotation is stored in the system history. Finally, the user can filter the annotations that characterize a specified area by a statistical filter that compares the annotation frequency with the neighborhood.

1 Introduction to MDR and Comparison with Related Works

The natural disasters require territory monitoring. SMAT is a distributed system for territory monitoring by means of Unmanned Aircraft Vehicles (UAVs). A UAV is equipped with payload sensors that download video of the target territory in the system.

The system should support the work of multiple operators. The operators can provide additional maps with annotations, metadata and accompanying files extracted from external sources (such as the web).

The software prototype that we describe in this paper is MetaData Retrieval (MDR). It is a geo-spatial web service integrated in the SOA architecture of the SMAT system. The main focus of MDR is to provide additional information on the locations included in cartographic maps, referred to as metadata. The availability of up-to-date cartographic maps is one of the main motivations of MDR because the maps get soon outdated. In addition, cartographic maps are often thematic and do not contain all the information that is needed by any user. On the contrary, there exists on the web a large amount of information on the geographical areas generated by Volunteered Graphic Information (VGI) projects by the everyday experience of people and by the integration of different cartographic databases: OpenStreetMap [3] (a free editable map of the whole world) and GeoNames [2] (the description and definition of over eight million named locations).

From this viewpoint, MDR is an innovative work because it integrates the information from VGI in a territory surveillance project by means of UAVs. For territory monitoring there is the need to retrieve and present the history of the annotations on a certain area of interest, to interactively explore the annotations by category, time, spatial object and compare them with the current status of the area. This exploratory interaction makes emerge easily the differences between the current geographical area and the past. This procedure is not present in VGI projects like LinkedGeoData [1] which work on the ontology for the annotations but do not provide a temporal view of the annotations in a selected area.

The information on locations is checked by MDR to be in a well-formed format and not redundant w.r.t. the history (i.e., containing no duplicates). In addition, annotations can be filtered by the user with a statistical filter that is based on the assumption of spatial auto-correlation. Significant annotations are those ones whose frequency in the area is an anomaly w.r.t. the frequency distribution in the neighborhood. In this way, it gives a characterization of the selected area and guarantees an increased level of reliability against noise and users errors. In this respect MDR is again a novelty w.r.t. the state of the art. For instance, [4] searches for frequent patterns in annotations but it does not filter annotations by outlier detection. [5] characterizes an area with a selection of the spatial features but it is much more difficult to tune because it considers three nested areas and user defined thresholds.

2   MDR Description

MDR is based on the annotations present in OpenStreetMap (OSM) and GeoNames. OSM maps are made up of three basic elements that form an ontological description: nodes, ways and relations. Each element has an arbitrary number of properties (tags) which are key-value pairs: 'key' represents a broad concept; 'value' is a specialization (like in key=historic; value=monument). A node represents a map feature or a standalone entity: it consists of the latitude and longitude of the location in the geocentric coordinate system, the user name who provided the data, a time-stamp and of additional feature attributes specified by key-value pairs.

GeoNames provides location descriptions which consist in the official name, a general description, demographic data, images, etc. Similarly to OSM, it provides a set of web services that provide an XML file with the annotations of the selected area.

Figure 1 is a screen-shot of MDR with its main output: an annotated, interactive cartographic map. The left-hand side of the window displays a tree-like arrangement that helps the user to index and browse the annotations. They are ordered first by location, then by retrieval time and finally by category (key). Each annotation on the left corresponds to an icon that can be correctly geo-referenced in the map on the right. The tree is navigable: the user can choose a specific spatial object, a point in time and a specific annotation category in the tree. As a consequence the relative annotations are correctly positioned on the map. If the user chooses a specific annotation from the list, a message box opens showing descriptive information on the location.
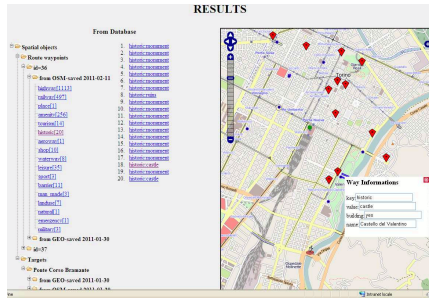
Fig. 1. Interactive map with annotations from the historical database
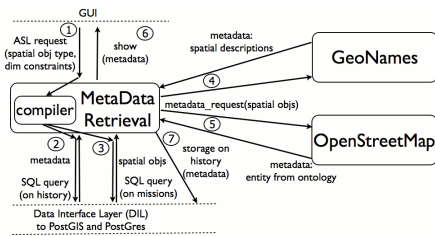

Fig. 2. Software architecture for MDR

2.1 Architecture Description

In order to accomplish the SMAT project goals, SOA is the suitable architectural choice because it allows the integration of different independent systems and services. Probably the best known spatial web services are Google Maps/Earth. In MDR, however, we chose to use open source software components (like Geoserver, OpenLayers for map visualization and development and PostGIS for the spatial database) that adopt open standards. Figure 2 shows the software architecture.

Any user's request to MDR searches for the metadata of some specified spatial objects. The spatial objects are involved in some of the facts contained in the system, general-purpose, multi-dimensional data model. In the case of the demo, the data model represents UAVs missions that are stored in the SMAT system database. The missions facts are described by the following, independent dimensions: UAV, sensor, target, airport, mission execution time, metadata time, space (spatial coordinates).

In MDR a user-friendly GUI allows the user to specify for which spatial objects involved in missions the annotations are requested. It is a sort of Query By Example which shows each dimension and allows the selection of any combination of dimensions values that allows the identification of the spatial objects of interest. The value selection is performed in a smart way because a drop-down list is presented with meaningful values. As regards the output annotations, the user can specify by a drop-down list the maximum distance (in Km) allowed between the spatial objects and the nearby locations whose annotations will be displayed. The dimension constraints generate a specification in an Abstract Specification Language (ASL) whose meaning is: *Retrieve the metadata associated to the specified spatial objects involved in the missions satisfying all the constraints*.

The MDR compiler translates the ASL query into a set of SQL queries for the database: the first type of SQL query retrieves the spatial objects satisfying the mission constraints and returns the spatial coordinates of the locations necessary to perform a web service request to OSM and GeoNames. The second query returns the metadata already stored in the system database for the same spatial objects. The complexity of the generation of a SQL query is linear in the number of constrained dimensions, being them independent as dimensions of a data warehouse star schema.

The answer of OSM and GeoNames is constituted by an XML file for each spatial object. Each file is parsed and the identified tags are presented in the output page grouped by category and ordered by time. Figure 1 presents the results for a query asking the annotations around the bridges over the river Po in Torino. Annotations show the historic monuments as well as other categories (railway, highway) and the categories corresponding to public services (like parkings, shops, amenities).

3   Geographic Characterization

We describe here how we obtain the content characterization of a geographical area. The characterization occurs in terms of the concepts corresponding to the tag categories provided as annotations. The filter consists in the extraction of the tags that are significant by a statistical validation method. It compares the frequency of each tag encountered in the given area, with the frequency distribution of the tags of the same category in the surrounding geographical areas.

We sketch the way in which we draw the sample from which we generated the frequency distribution. We build a regular grid composed of a total of 49 cells, surrounding the central area. All the cells of the grid have equal surface area of the central, target cell: thus in any cell each tag category has the same probability to appear. The frequency of each tag category appearing in the central cell is collected also in all the neighborhood areas: their distribution in the neighborhood is generated and compared with the frequency in the central cell. We perform a standard, statistical test on the frequency of each category with a very low significance level ($\alpha = 0.001$). The frequencies of the tag categories in the central area that are outliers of the frequency distribution in the surrounding areas are highlighted as the interesting ones because surprising. In fact, given the property of spatial auto-correlation of the features, most of the tag categories are expected to occur also in the neighborhood with a frequency similar to the central cell. The tags that are selected are: (i) the tags on which the majority of the users agree; (ii) the tags that characterize the area because discriminate it with the surroundings. As an example for the central area of Torino, significant tags result the fountains, the restaurants, the tram stops, the bicycle rental sites which confirm the characterization of the touristic area in contrast with the residential neighbourhood.

References

1. Auer, S., Dietzold, S., Jens, L., Hellmann, S., Aumueller, D.: Triplify lightweight linked data publication from relational databases. In: Proc. of WWW (2009)
2. GeoNames, `http://www.geonames.org` (Retrieved 10-11-2010)
3. Haklay, M., Weber, P.: Openstreetmap: User-generated street map. IEEE Pervasive Computing 7(4), 12–18 (2008)
4. Sengstock, C., Gertz, M.: Exploring volunteered geographic information using scale-dependent frequent pattern mining. In: Proceedings of GIScience (2010)
5. Tomko, M., Pulves, R.: Venice, city of canals: Characterizing regions through content classification. Transactions in GIS 7, 295–314 (2009)