Chapter 17
Geographical Map Annotation with Significant Tags Available from Social Networks

Elena Roglia
University of Turin, Italy

Rosa Meo
University of Turin, Italy

Enrico Ponassi
University of Turin, Italy

ABSTRACT
In this chapter we describe how to extract relevant information on a geographical area from information that users share and provide by means of their mobiles or personal digital assistants, thanks to Web 2.0 applications such as OpenStreetMap, Geonames, Flickr, and GoogleMaps. These Web 2.0 applications represent, store, and process information in an XML format. We analyze and use this information to enrich the content of the cartographic map of a given geographical area with up-to-date information. In addition we provide a characterization of the map by selection of the annotations that differentiate the given map from the surrounding areas. This occurs by means of statistical tests on the annotations frequency in the different geographical areas. We present the results of an experimental section in which we show that the content characterization is meaningful, statistically significant, and usefully concise.

INTRODUCTION
Today most of the data available for spatial data mining can be found on the Web. Web 2.0 technologies enable users to add information to Internet pages, allowing a two-way flow of information, from the producer to users, and vice versa (O’Reilly, 2010). Thus, the users and producers, whether as experts or amateurs, were transformed to producers of geo-data (Budhathoki, Bruce, & Nedovic-Budic, 2008). Wikipedia, the first major application of Web 2.0, together with Geonames, Flickr, OpenStreetMap and other blogs and social networks are freely available tools that provide
information coming from the community of their users on the geographical domain. Information is updated frequently and can be shared as a free source of information (Wikipedia; Geonames; Flickr; OpenStreetMap). In addition, most of the mentioned sites provide API - Application Programming Interfaces - to end users. These sets of procedures allow programmers to retrieve and manipulate data in different formats, including XML (W3C XML), KML (OGC KML) or other XML-based formats, as in the case of OpenStreetMap, which offers its own data format: OSM. This type of data representation enabled the scientific community to integrate traditional spatial data mining techniques with XML mining techniques.

In this chapter we focus on the content characterization of geographical areas using OpenStreetMap tag elements. This need arises from two observations.

1. The cartographic maps describing the territory are rich of detailed information, but are costly and get outdated very soon. For instance, unfortunately, in Italy there are only a few regional departments that maintain an updated and detailed cartography of their region. Many other regions have an outdated cartography that does not respond anymore to the users’ needs. Furthermore, a cartographic map is often thematic and does not contain all the information that is needed by any user. For instance, there are maps devoted to tourists, maps of trasportation services, maps for military activities, etc.

2. On the Internet, there exists very often a large amount of information on the geographical areas generated by the everyday experience of people. People provide geographical information through handhelds, pocket PCs or mobile phones connected to Internet while traveling or simply while moving. This information is frequently updated by the active users of social networks. The idea to enrich the cartography with these fresh annotations arises immediately.

Information is generally inserted by users of social networks for free and are not always strictly controlled by a system moderator. Suppliers are allowed to annotate any location or spatial object they wish by associating it with a tag. OpenStreetMap helps the users in annotating locations by providing them an ontology of spatial objects. For each object class a tag is provided. Tags have the form of a key:value pair, such as historic:monument or historic:castle in which the key (historic, in the example) represents the object broad concept and the value (monument or castle) specializes it. The annotation system also allows the specification of additional information that the user considers useful. For instance, with an annotation like amenity:restaurant a user might provide an additional attribute: cuisine:Chinese. Usually, users provide annotations on commodities (like a bike shop or a mountain trail), on general interest locations (like airports or tourists’ attractions) or on spatial objects by their purpose (hospital or zoo).

This chapter describes how we obtain a content characterization of a geographical area. The characterization occurs in terms of the concepts corresponding to the tags provided as annotation by the users on that area. A problem might arise in this process if a big number of tags might be provided by users, especially in certain metropolitan areas. In addition, some of the tags could not be relevant or interesting or be the result of a mistake. We look at this misleading result like the effect of the superimposition of noise on the valuable information.

In order to eliminate the noisy effects and validate the users annotation process we propose to apply a filter to the tags. The filter consists in the extraction of the tags that appear to be significant by a statistical validation method. This validation method compares the frequency of occurrence of each tag encountered in the given area, with the