Geospatial Data Mining on the Web: Discovering Locations of Emergency Service Facilities

Wenwen Li[1], Michael F. Goodchild[2], Richard L. Church[2], and Bin Zhou[3]

[1] GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe AZ 85287
Wenwen@asu.edu
[2] Department of Geography, University of California, Santa Barbara
Santa Barbara, CA 93106
{good,church}@geog.ucsb.edu
[3] Institute of Oceanographic Instrumentation, Shandong Academy of Sciences
Qingdao, Shandong, China 266001
senosy@gmail.com

Abstract. Identifying location-based information from the WWW, such as street addresses of emergency service facilities, has become increasingly popular. However, current Web-mining tools such as Google's crawler are designed to index webpages on the Internet instead of considering location information with a smaller granularity as an indexable object. This always leads to low recall of the search results. In order to retrieve the location-based information on the ever-expanding Internet with almost-unstructured Web data, there is a need of an effective Web-mining mechanism that is capable of extracting desired spatial data on the right webpages within the right scope. In this paper, we report our efforts towards automated location-information retrieval by developing a knowledge-based Web mining tool, CyberMiner, that adopts (1) a geospatial taxonomy to determine the starting URLs and domains for the spatial Web mining, (2) a rule-based forward and backward screening algorithm for efficient address extraction, and (3) inductive-learning-based semantic analysis to discover patterns of street addresses of interest. The retrieval of locations of all fire stations within Los Angeles County, California is used as a case study.

Keywords: Emergency service facilities, Web data mining, information extraction, information retrieval, ontology, inductive learning, location-based services.

1 Introduction

Although it has only been 22 years since its advent, the World Wide Web (WWW) has significantly changed the way that information is shared, delivered, and discovered. Recently, a new wave of technological innovation - the emergence of Web 2.0 and Web 3.0, such as social networks, government surveillance and citizen sensors [5] - has led to an explosion of Web content, and brought us into the era of Big Data [18]. Statistical reports show that by 2008 the amount of data on the Internet had reached 500bn gigabytes [21], and the indexed Web contained at least 11.5 billion pages [6].

S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 552–563, 2012.

This information explosion on the Web poses tremendous challenges for various information-retrieval tasks [12].

Within this massive amount of data, identifying location-based information, such as street address and place name, has become very popular, due to the desire to map this information from cyberspace to the physical world. As a type of spatial data, locations of emergency service facilities are especially important in protecting people's lives and safety, and for government agencies to provide real-time emergency response. Taking fire stations as an example, besides the aforementioned functions, insurance companies need the locations of all fire stations within and near a community to determine the insurance costs to be paid by a household. Decision-makers long for this location information to obtain the urban footprint of each fire station and to plan the optimal placement of fire stations within a region.

Presently, most Internet uses obtain WWW information from search engines [16]. However, the commercial search engines such as Google are designed to index webpages on the Internet instead of considering location information that has smaller granularity as an indexable object. Therefore, these search engines always lead to a low recall rate in search results. Fig.1 shows the search results for fire stations within the city of Santa Barbara, CA, from Google. Red pinpoints are the Googled results and blue pinpoints are the actual locations of fire stations within that city. It can be seen that except for *C* (to the west of the green arrow) overlapping with its actual location (blue pinpoint), all of the results are irrelevant.

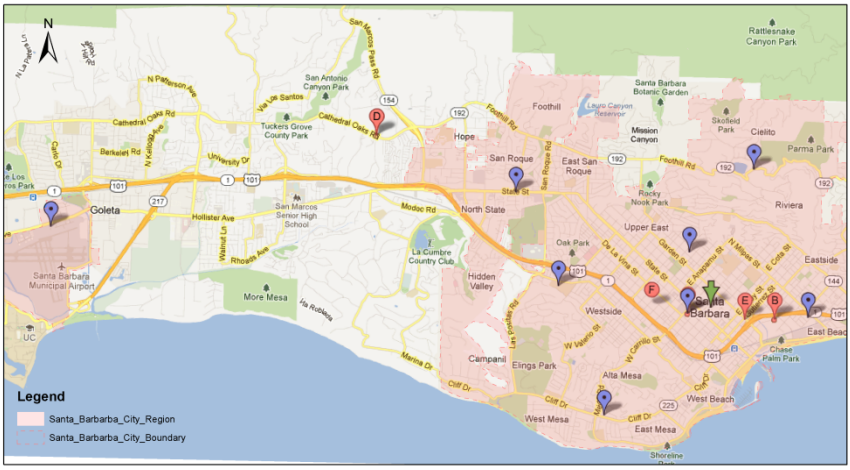

Fig. 1. Results of a Google search for locations of fire stations in the city of Santa Barbara, CA. The pink region shows the geographic extent of Santa Barbara.

In order to retrieve location-based information on the ever-expanding Internet and its almost-unstructured Web data, there is a need of an effective Web-mining mechanism that is capable of extracting desired data on the right webpages within the right scope. In this paper, we report our efforts in developing a knowledge-based Web-mining tool, CyberMiner, that adopts geospatial ontology, a forward- and backward-screening algorithm, and inductive learning for automated location information