Algorithms and Data Structures for Large Scale Geographic Information Systems

João Valença
University of Coimbra
Coimbra, Portugal
valenca@student.dei.uc.pt

Luís Paquete
University of Coimbra
Coimbra, Portugal
paquete@dei.uc.pt

Pedro Reino
Smartgeo Solutions
Lisboa, Portugal
pedro.reino@smartgeo.pt

Carlos Caçador
Smartgeo Solutions
Lisboa, Portugal
carlos.cacador@smartgeo.pt

Abstract

Currently, there is a need to process a large amount of geographic data before it can be managed and analysed. This project aims to develop a GIS application operating through a Web platform in order to allow for a low cost and simplified integration, management and manipulation of georeferenced information.

This project aims to develop a way to efficiently extract a subset of a collection of geographic points whilst keeping the representativeness as close as possible to the whole set. Special emphasis is given to the implementation of efficient clustering algorithms for finding a representative set of points in a map, which can be recast as a *k-center* problem. The approaches covered in this work include exact algorithms for finding minimum coverage subsets, as well as heuristic approaches for finding good approximations in real time.

Keywords: Geographic Clustering, k-Center, Coverage, Branch-and-Bound, Delaunay Triangulations

1 Objective

One obstacle when representing large amounts of geographic data is that the sheer volume of points to display can be overwhelming for a human, as well as computationally intensive to render for a machine. As such, there is a need to develop and implement a viable way to reliably calculate and display a subset of geographic points, whilst keeping a degree of representability of the larger set, so that as little information as possible is absent when the representative subset is shown.

The purpose of this project is to develop a real-time algorithm that can analyse geographic data provided by a geographic information system infrastructure developed and maintained at Smartgeo. More precisely, the developed algorithm has to be able to aggregate and select geographic points according to a given a set of criteria in real time. The chosen subset should keep a measure of representativeness of the larger set.

2 Problem Definition

Representativeness consists of finding a subset of points in a larger set. The subset chosen should be able to keep some specified properties of the original set, such as density, or general distribution. As such there can be many ways to define representativeness.

One such way is finding the subset that minimises the maximum distance between the points not chosen, and their closest counterparts within the chosen set. This notion of representativeness is known in the optimisation field as coverage, and finding the subset of k points that minimises the value of coverage is known as the *k-center* problem [3].

In order to address this problem, a few algorithms were developed and implemented. Optimal approaches to this problem include integer linear programming, as well as incremental branch-and-bound algorithms based around geometric properties of the points, such as Delaunay triangulations [2] and Hilbert curves [4].

However, the inherent overhead in these algorithms makes them unsuitable for use in practical applications. For achieving an acceptable suboptimal solution within a reasonable time frame of what is expected in a real-time web

application, heuristic methods need to be used. The algorithms developed have to be benchmarked for quality and performance in order to chose the most fitting one for integration with the web platform.

3 Architecture

The application will display a rectangular window, showing a cut of geographical region containing a set of points. The algorithm chosen will need to be able to choose a representative set of points within the cut quickly, as well as be able to recalculate a new set points for a new cut, resulting from panning or zooming the display window over the region.

The algorithms tested include exact algorithms for finding the minimum coverage subset. These are comprised of two branch-and-bound approaches: a naïve incremental approach and a geometric incremental approach that makes use of the properties of Delaunay triangulations in order to speed up point location queries via the Greedy Routing algorithm [1]. An exact integer linear programming approach is benchmarked as well. This project also includes a few heuristic approaches and approximation algorithm techniques to solve the problem more efficiently.

The algorithm will serve as the middleware responsible for filtering the response of a GIS server to a Web Feature Service, or WFS request. WFS lists the geographic coordinates of the points to be represented in an image by the coordinates mapped into an orthogonal plane, representing an image containing the cut of the region requested by the application.

Finally, a web application will receive the response filtered by the best algorithm and interacted with by a human user.

The candidate algorithms will be tested and benchmarked using data from the Open Street Map project. The project features large quantities of open source geographic data, as well as a versatile API for fetching data in the WFS standard.

References

[1] P. Bose and P. Morin. Online routing in triangulations. In *Algorithms and Computation*, pages 113–122. Springer, 1999.

[2] M. De Berg, M. Van Kreveld, M. Overmars, and O. C. Schwarzkopf. *Computational Geometry*. Springer, 2000.

[3] N. Megiddo and A. Tamir. New results on the complexity of p-centre problems. *SIAM Journal on Computing*, 12(4):751–758, 1983.

[4] H. Sagan. *Space-Filling Curves*, volume 18. Springer-Verlag New York, 1994.