M.Sc. Johannes Kröger
HafenCity University Hamburg, Lab for Geoinformatics and Geovisualization (g2lab)
Überseeallee 16, 20457 Hamburg
johannes.kroeger@hcu-hamburg.de

Significant locations in ancillary data as seeds for typical use cases of cartographic point clustering

Abstract

To avoid data crowding, overlapping of marker icons or to minimize load on the client's computer, clustering methods are often used on point data in interactive maps. Here, instead of displaying a marker icon for each point, groups of them are aggregated into clusters and their display is limited to one marker icon for the whole group. Popular web mapping libraries and services offer point clustering as a base feature, easily enabled by users without requiring any additional knowledge.

Usually either greedy or grid-based clustering algorithms are used due to their low computational cost and general applicability. In greedy clustering randomly chosen points seed clusters that aggregate their neighbors until all points are assigned to a cluster. In grid-based clustering a polygonal grid, often made out of squares, is placed over the area of interest and points are aggregated per cell.

We argue that the resulting cluster distributions of these standard algorithms often mislead, at least in maps where the spatial distribution of the points follows an underlying pattern like population density. A typical store locator map for example, commonly displayed on company websites, should allow potential customers to determine if any store is available at a certain location. Here such clustering can drastically change the validity of the map: A group of stores, located naturally in a city, might be torn apart into separate clusters, leaving the city displayed as uncovered by any cluster. The city might sit just between the randomly chosen cluster centers of a greedy clustering approach or right on the border between cells in a grid-based clustering. If this happens, the map failed its purpose of showing that stores exist in the city.

We work on a new, straightforward clustering method, where precalculated and weighted points at local maxima of e.g. population density serve as seeds for the clustering process. As the seeds serve as “context-aware” cluster anchors for data that is, or at least appears to be, closely correlated to population density, our algorithm is applicable for data of such nature. The

concept itself is applicable for other phenomena where previous knowledge exists.

In our approach the cluster seeding points are located on the local population maxima. These local maxima can be calculated on varying scales using grid-based population maps or derived from datasets of populated places like settlements or metropolitan areas. Each seed's weight is used to calculate the extends of its catchment area. Neighboring points of the dataset-to-be-clustered are then aggregated into clusters per catchment area. Each seed's weight and its influence on the catchment area can be specified dependent on the scale, allowing a granular and dynamic control over suitable locations.

Detailed results are yet to be evaluated but prototyping and preliminary tests suggest that this approach can improve the quality of point clustering for specific use cases. Both analysis of quantitative metrics as well as a study on user acceptance and task efficiency will be performed. Further work may include a free, global, multi-level dataset of suitable cluster seeding points for these purposes.

Main problems at the current stage of research are determining an appropriate definition of the ancillary data's local maxima (or similar) depending on the scale and extends of the map, and handling screen space versus world space in terms of the map data and map display.

I am at a very early stage of this research and could use brutally honest feedback and constructive ideas/tips.