UDC 004.9+004.021

GEOLOCATION-RELATED DATA CLUSTERING METHODS

ARTUR DRYOMOV, ARKADY OSKIN
Polotsk State University, Belarus

Displaying various points on a map it is possible to get high points density in certain areas of the map. This kind of behaviour is not desired for an end user of the map, mostly because it is not trivial for a person to manipulate points in such environments. The issue can be solved by grouping points using clustering methods.

Maps in all their variety are essential for proper interpretation and representation of overwhelming amount of geographical data. This kind of data surrounds us every day in any form imaginable. One of the first things we do in the morning is the commute from home to work and, as a consequence, vice versa action in the evening. This can sound like a common sense, but it requires deep geolocation-related information knowledge about at least two points: source and destination. Fortunately enough with the current informational technology integration level in our everyday lives we have more or less reliable online mapping solutions like Google Maps, Yandex Maps, Apple Maps and so on, all of them are quite popular. Their users use them a they please to perform various tasks: search for a location and a route to that location, view location details and reviews, plan complex routes for a set of locations and much more than that. Soon enough a complex issue arises — what to do if there is a large amount of locations needed to be displayed on a map. The straightforward solution is just to put them as it is without any preprocessing. The final result can be kind of disappointing — a lot of markers, associated with locations, on a limited in size will look like a tightly filled with markers field without any option to take a look or select a single one among them, visually overloading the map and its interaction. There are some solutions to deal with such sort of situations.

Clustering or Cluster analysis is a specific operation of transforming a set of various objects into groups. Resulting groups can be characterized and are different one from another. Grouped objects in a single picked group are more similar to one another than other ones in different groups. These groups of objects are called as clusters. The clustering task is one of primary areas of exploratory data mining. The operation itself has a wide appliance: form machine learning and image analysis to data compression and computer graphics. It is important to notice that clustering is not an algorithm, but a general task. Analysis and cluster grouping can be performed using various methods, ways and notions about what constitutes a cluster itself and how to differentiate one cluster from another. Most known notions are distance-related metrics, dense areas of data space, statistical distributions or intervals. This way clustering can be characterized as a multi-objective optimization issue. Cluster analysis is not a fixed-steps issue for most of the data, it is an iterative process of trial and failure involved in the knowledge discovery. Sometimes it is necessary to change algorithms and its parameters until the result is available. Depending on specific data set and desired results form distance function, density threshold and number of resulting clusters can and should be changed in the process of analysis itself [1].

Geomapping data, as it was noticed before, requires clustering when working with a large amount of location-related data. Most of all the issue relates to map markers — map objects associated with specific locations. Such objects can be represented as pins, location dots and signs. The issue of representing an overwhelming amount of pins on a fixed map area will lead to a field filled with pins without any underlying context. At the same time, another digital map-related parameter should be considered called zoom. While observing a map area using maximum and minimum zoom results will be different. For example, when placing a million pins on a single street, maximum zoom (a house-level one) will lead to observing a spread of pins, while minimum zoom (a planetary-level one) will show a dot of highly dense area of pins.

The proposed solution is to use clustering to group pins based on their location and current zoom level. The desired behaviour will replace pins with pin clusters and vice versa while zooming out and in respectively. The proper implementation will use clustering to optimize displayed pins count to exclude results with overwhelming data displayed on the map [2].

There are two solutions related to geographical-related data clustering.

- Square based clustering.
- Distance based clustering.

Square based clustering, as the name says, does the clustering via dividing a map to squares. The square size depends on a current map zoom level. Map markers appearing in the calculated square are grouped into a cluster. The technique is quite simple.

1. Choose a coordinate system on a map and split it to squares of a chosen size.
2. Start iterating over all locations available in a set.
   a. Find out which square hosts the location.
   b. Put location into a square-related subset.
3. Finish iterating over all locations available in the set.
4. Start iterating over all resulting square subsets.
   a. Calculate a centre of the square.
   b. Put a cluster marker into a calculated centre.
5. Finish iterating over all resulting square subsets.

The similar approach is based on administrative units clustering. This way a map is not divided into squares and markers are tested against fitting into specific predetermined geographical regions, most likely being equivalent to real units of the selected country [3].

The technique has some limitations.

– Because splitting a map to squares is mostly artificial the resulting markers set looks artificial as well, basically being a perfectly aligned markers on the same size one from another. In this way the geographical data context can be lost.

– Two or more markers can be very close to one another but be in separate squares. As a result, clustering will not work properly.

Distance based clustering is a smarter but less efficient way of grouping markers into clusters. Instead of splitting the map into squares or regions markers are grouped by their respective distance to one another. For example, it is possible to combine locations into a cluster inside 10 kilometres radius. There is one issue though. Kilometres, meters and such metric units have different meaning based on the current zoom level. A zoomed in map displays a distance unit differently than a zoomed out map. This needs to be considered while implementing the algorithm. Even more — the clustering operation should be redone on each zoom level change [4].

1. Choose a random location from a location set.
2. Form a current cluster.
3. Remove the chosen location from the location set.
4. Start iterating over all locations available in the set.
   a. If a current location fits the desired distance criteria regarding the chosen location:
   b. include it into a current cluster;
   c. remove it from a location set.
5. Finish iterating over all locations available in a set.
6. Repeat from the beginning until the set is empty.
7. Start iterating over resulting clusters.
   a. Calculate a geographical centre of locations in the cluster.
   b. Put a cluster marker into a calculated centre.
8. Finish iterating over resulting clusters.

This approach gives better results, but is not as efficient as the square-based clustering — it is required to pass through all locations multiple times instead of a single pass with the square-based method.

There are other, more complex, methods, that can be researched and benchmarked in the future developments, but it clearly seems that the issue is real and has practical applications in the current state of the world.

REFERENCES

1. Dudek, A. Dynamic Classification of Geographic Points on Google Maps / A. Dudek ; University of Lodz. – Lodz, 2014. — 10 p.
2. Meert, W. Clustering Maps / W. Meert ; Katholieke Universiteit Leuven. — Leuven, 2006. — 62 p.
3. Hot, E. Soil data clustering / E. Hot, V. Popovic-Bugarin // Telfor Journal. — 2016. — 6 p.
4. Chen, X. Efficient Filtering and Clustering Mechanism forGoogle Maps / X. Chen // Journal of Advanced Management Science. — 2013. — 5 p.