FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Diplomarbeitspräsentation

Geocluster: Server-side clustering for mapping in Drupal based on Geohash

Josef Dabernig

Technische Universität Wien
Institut für Softwaretechnik und Interaktive Systeme
Arbeitsbereich: Information & Software Engineering Group
Betreuer: O.Univ.Prof. Dr. A Min Tjoa

Masterstudium:
Software Engineering
& Internet Computing

Problem

Maps visualize data in an intuitive way. Performance and readability of digital mapping applications decreases when displaying large amounts of data. Client-side clustering uses JavaScript to group overlapping items. Server-side clustering is needed when too many items slow down processing and create network bottle necks.

Goals

- Implement real-time, server-side clustering
- Cluster up to 1,000,000 items within 1 second
- Visualize clusters on an interactive map
- Integrate with the Drupal framework
- Publish under the Open Source GPL license
- Implement use cases and evaluate results

Approach

- Research clustering, mapping and visualization
- Evaluate state-of-the-art technologies
- Design a scalable algorithm for clustering
- Implement and test the algorithm

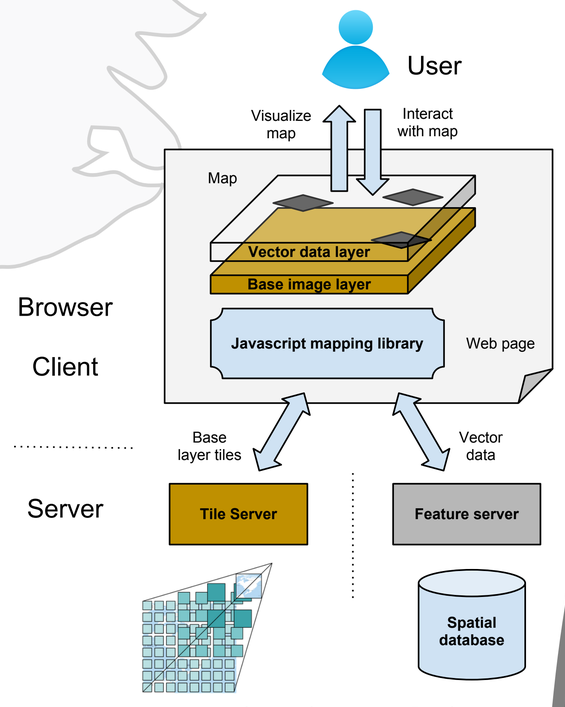
Geohash space decomposition on level 1. The letter „D" covers parts of the Americas

Clustering is the task of grouping unlabeled data in an automated way. The thesis researches cluster analysis to create an algorithm for server-side clustering with maps.

Geohash is a latitude/longitude geocode system based on the Morton order. Coordinates are encoded as string identifiers with a hierarchical spatial structure.

Algorithm considerations

- *Pattern representation*: spatial clusters
- *Proximity measure*: Euclidean distance
- *Cluster type*: prototype-based
- *Algorithm*: based on Geohash



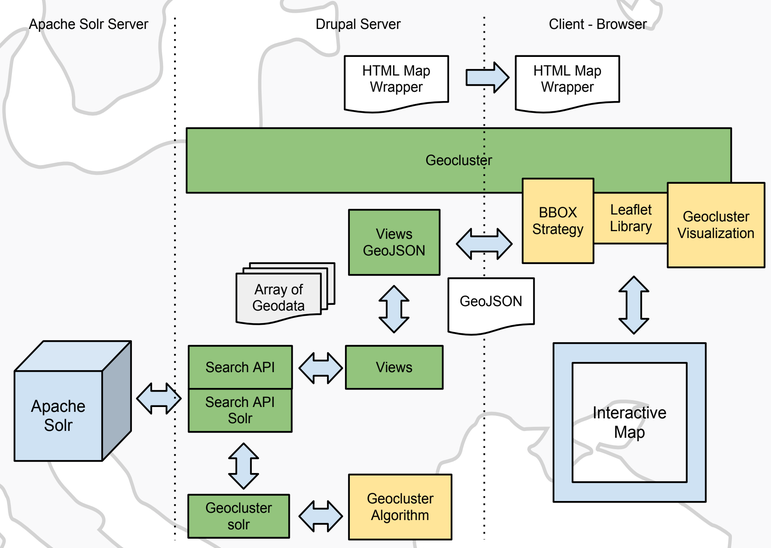
A modern web mapping stack

Clustering

Implementation

Create a Geohash-based hierarchical spatial index
1) initialize algorithm variables (cluster level)
2) pre-cluster points based on Geohash
3) merge clusters by neighbor-check

The algorithm has been integrated into the Drupal mapping stack as shown in the figure below:

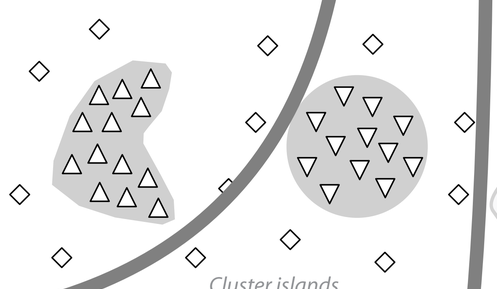


Geocluster Solr architecture overview

Mapping

- Spatial data is represented by points, lines or polygons in vector format or rastered images
- Projections map the geoid earth onto a planar surface which causes distortion
- A modern web mapping stack uses image base tiles with overlays of vector data
- The slippy map is rendered client-side by a JavaScript mapping library

The Drupal mapping stack has been studied for integration for a server-side clustering solution.



Cluster islands

Geocluster

Visualization

Foundations of geovisualization, visual variables, data exploration techniques and clutter reduction have been researched. A state-of-the-art analysis enumerates map visualization types and techniques for putting clustered, multi-variate data on maps.

- *Map types*: Geographic maps with markers, Heat/choropleth maps, Dot grid maps and Voronoi maps
- *Cluster visualization techniques*: Icon-based/Glyphs, Pixel-oriented as well as Geometric techniques and Diagrams.

An evaluation classifies the stated techniques for cluster visualization on maps, based on exploratory analysis.

Simple glyph types

Drupal

Drupal is a free and open source content management system and framework. Developed and maintained by an international community, it currently backs more than 2% of all websites.

The Drupal mapping stack has been evaluated for integration of a server-side clustering implementation, including modules for spatial data storage and presentation.

Geocluster integrates with state-of-the-art Drupal 7 modules like Geofield, Views, Leaflet to provide interactive, scalable, clustered maps.

It has been released under the GPL license and can be downloaded from:

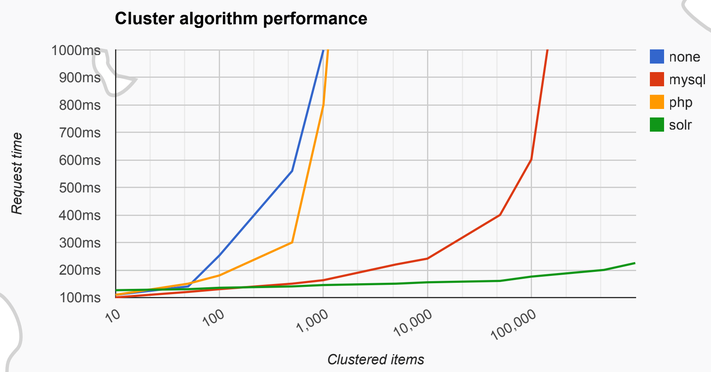
http://drupal.org/project/geocluster

Results

Two use cases have been realized and evaluated for performance and visualization: a *geocluster demo use case* and a *GeoRecruiter prototype* that extends the Recruiter distribution for job boards in Drupal 7.

The **performance tests** show that one of the 3 algorithm implementations fulfills the objective:

- the PHP implementation doesn't scale well
- the MySQL clustering scales up to 100,000 items
- the Solr version scales beyond 1,000,000 items



Geocluster performance

Kontakt: http://dasjo.at