Diploma Thesis – Exposé - Draft

Server-side geo clustering in Drupal based on Geohash

by Josef Dabernig e0927232@student.tuwien.ac.at

Introduction

Digital mapping applications on the Internet are strongly emerging. Big players like Google Maps[1] and OpenStreetMap[2] provide online maps, that users can view and interact with.

Maps allow telling stories and communicating data in a visual way. Using open source tools such as TileMill[3] and online services like CloudMade[4], more and more people are able to create their own custom maps. Content management systems and frameworks like Drupal[5] provide tools to add, edit and visualize geographic data on maps. This allows to integrate interactive map applications into web sites.

Clustering[6] is a technique in Data Mining[7] for grouping objects with similarities. In geospatial map applications, clustering by distance is a common task. By preventing overlap of symbols, it enhances readability of maps. In addition to that, clustering is used to improve performance of data-heavy maps.

Geohash[8] encodes geographic location data as string with a hierarchical spatial data structure. In theory, these attributes allow for implementing a performance-optimized clustering algorithm for geospatial data. Drupal is an Open Source Web Framework with more than 18,000 extension modules available.

This thesis therefore focuses on implementing an algorithm for server-side geo clustering in Drupal based on Geohash.

Problem statement

The multitude of available programs and tools for leveraging online maps with Drupal is developing quickly. There exist various modules like Geofield[9], OpenLayers[10], Leaflet[11] and Geocoder[12] which allow to integrate interactive, data-driven maps in websites built with Drupal. One missing gap in that field is still the possibility to visualize thousands of points on a map in a performant and usable way.

Various Javascript libraries like Google MarkerCluster[13] exist to cluster geo data on the client-side. This enhances performance and readability of data-heavy map applications. But still, all data needs to be transferred to the client and processed on a potentially slower end user device.

Clustering data on the server-side shifts the load from the client and allows to display many points in a

1 https://maps.google.at/
2 http://www.openstreetmap.org/
3 http://mapbox.com/tilemill/
4 http://cloudmade.com/
5 http://drupal.org/
6 http://en.wikipedia.org/wiki/Cluster_analysis
7 http://en.wikipedia.org/wiki/Data_mining
8 http://en.wikipedia.org/wiki/Geohash
9 http://drupal.org/project/geofield
10 http://drupal.org/project/openlayers
11 http://drupal.org/project/leaflet
12 http://drupal.org/project/geocoder
13 https://developers.google.com/maps/articles/toomanymarkers#markerclusterer

performant way. Professional services like maptimize[14] provide such a functionality, while in the open source space little libraries and frameworks exist for server-side clustering of geospatial data.

It is a common practice in PHP[15]-driven Drupal to leverage Apache Solr[16] for performant queries on large datasets. Therefore, a clustering solution that integrates well with Drupal and Apache Solr on the server-side is required.

Aim of the work

This thesis evaluates and compares different techniques in the field of web cartography which allow to provide clustered data-intense and interactive map applications using Drupal. Theoretic foundations for spatial clustering have to be evaluated, including different clustering approaches and their algorithms. Existing implementations available within the open source space should be reviewed and considered for incorporation or derivation.

The practical part of the thesis involves implementing a working prototype for server-side geo clustering based on Drupal 7. Integration with best-practice Drupal extension modules is one of the main targets: allow creating custom implementations based on widely used Drupal tools like Views[17].

Planned deliverables

- Algorithm for clustering geospatial data on the server-side.
- Clustering implementation as drop-in module for Drupal.
- Clustering implementation as Apache Solr plugin with Drupal integration.
- Visualization and interaction for clusters on the client-side.

Methodological Approach & Structure of the Work

This thesis aims at engaging with active contributors of Open Source communities like Drupal and in general in the geospatial web to incorporate state-of-the art technologies.

Approaches to the thesis will include

- researching open source mapping technologies
- evaluating different algorithms, libraries and tools
- engage with open source and open data communities to capture recent developments which haven't been published in a scientific way, yet

As part of the research for the thesis, during Computer Science Practical a prototype for open source data mapping in Austria has been developed: AustroFeedr[18] creates a universal & open source prototype for aggregation, processing and data publishing in real-time based on the PubSubHubbub-protocol. Created content can be displayed visually on maps and diagrams. By implementing the „AustroFeedr Hochwasser“ (flooding) real world use case in cooperation with the Austrian government, it showcases the technology as a monitoring system of river water levels for flood protection in Austria.

14 http://www.maptimize.com/
15 http://www.php.net/
16 http://drupal.org/project/search_api_solr
17 http://drupal.org/project/views
18 http://austrofeedr.at/

Mapping Foundations

The layered mapping stack

Online map applications are built of different components

- data sources
- map base layer
- data overlays with interaction

Parts of the map production process happens on the server side (e.g. base layer rendering) while further steps are computing on the client side (within the browser, e.g. dynamic display of data overlays).

Functional mapping metrics

It is planned to investigate current approaches for building customized maps on a functional level by the following metrics:

- **data**: which kinds of data drive the maps (linear, vector, ...)
- **base layers**: which map base layers exist (google maps, open street maps, ...)
- **data layers**: how is the data being represented on top (markers, symbols, charts, ...)
- **interaction**: which kinds of user interaction is supported (popups, panning, zooming, ...)

Technological mapping metrics

Evaluation of mapping tools will include a comparison under the following aspects:

- **data connectivity**: how to integrate different kinds of data with mapping technologies (e.g. csv imports, web services, …)
- **static vs dynamic**: how and why are different parts of maps realized using different techniques (e.g. base layers consist of static images, data overlay layers are rendered dynamically and provide interactivity)
- **interaction**: how is user interaction realized in the browser (e.g. javascript behaviors, …)

Technologies

- Representing geospatial data: Latitude/Longitude[19], WKT and WKB[20], GeoJSON[21], Geohash[22]
- Storing geospatial data: MySQL Spatial[23], PostGIS PostreSQL[24], Apache Solr LatLonType[25]
- Working with geospatial data: GEOS[26], Apache Solr Spatial Search[27]

19 http://en.wikipedia.org/wiki/Geographic_coordinate_system
20 http://en.wikipedia.org/wiki/Well-known_text
21 http://www.geojson.org/
22 http://en.wikipedia.org/wiki/Geohash
23 http://dev.mysql.com/doc/refman/5.0/en/spatial-extensions.html
24 http://postgis.refractions.net/
25 http://wiki.apache.org/solr/SpatialSearch#LatLonType
26 http://trac.osgeo.org/geos/
27 http://wiki.apache.org/solr/SpatialSearch

Drupal mapping foundations

- The Drupal Geo Stack (Best-practice implementations & modules)
- Storage: Geofield, PostGIS, Location
- Process: geoPHP
- Query and Display: OpenLayers, Leaflet, Views GeoJSON, Mapping

Clustering foundations

- Explain the field of Geospatial clustering, also named Analytical regionalization or Spatially constrained clustering.
- Compare different approaches
  - where to cluster (client-side, in the business logic, on the database level)
  - when to cluster (on-request or pre-calculate clusters and store them in the database)
- Algorithms & techniques
  - Well-separated, Prototype-based, Graph-based, Grid-based, Density-based and conceptual
  - Examples: K-Means, DBSCAN, Quadtrees, CORSO, Geohash
- Implementations
  - Client-side: Google MarkerClustering, MapBox clustr, Leaflet.markercluster
  - Server-side: clusterPy, Vizmo, Geoclustering for Drupal 6
- Visualization & interaction techniques

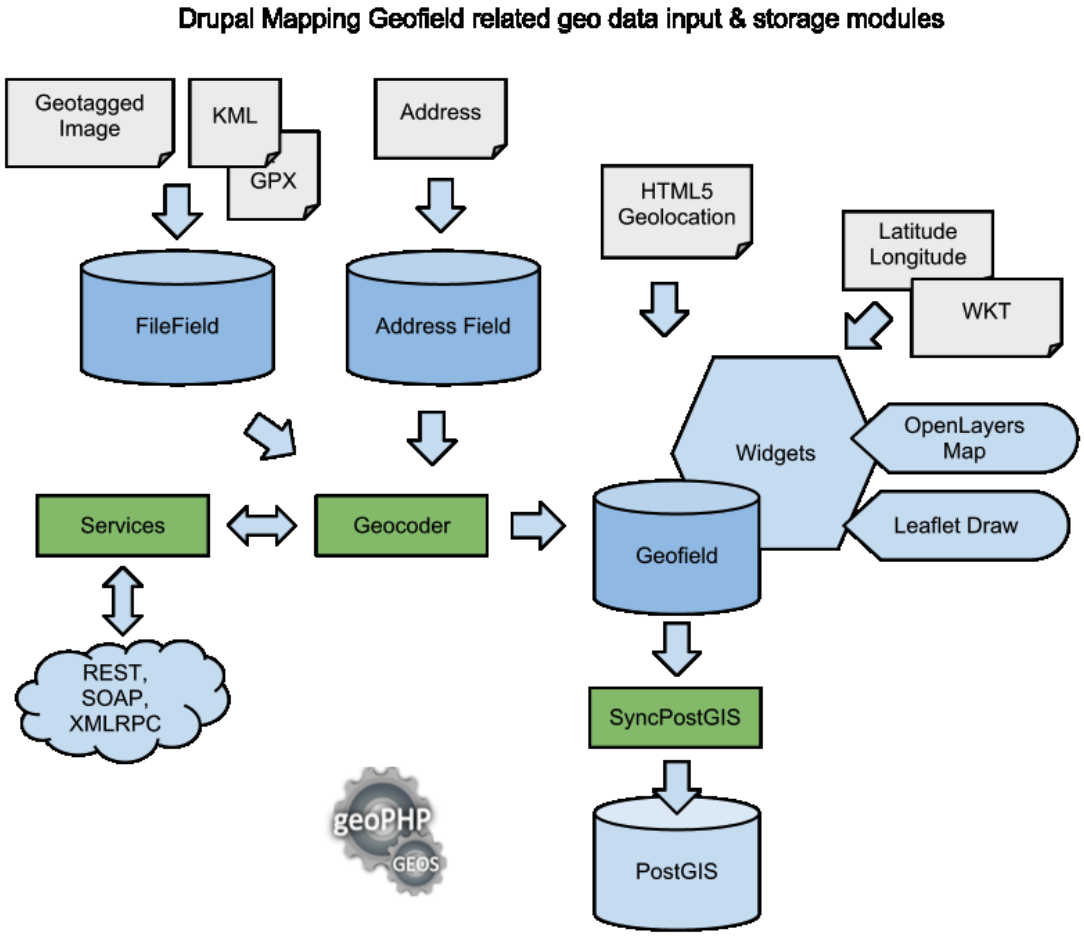
Use cases

- AustroFeedr: visualize flood incidents in Austria on a map, based on open data
- GeoRecruiter: find jobs on a map, based on the open source Drupal 7 distribution Recruiter

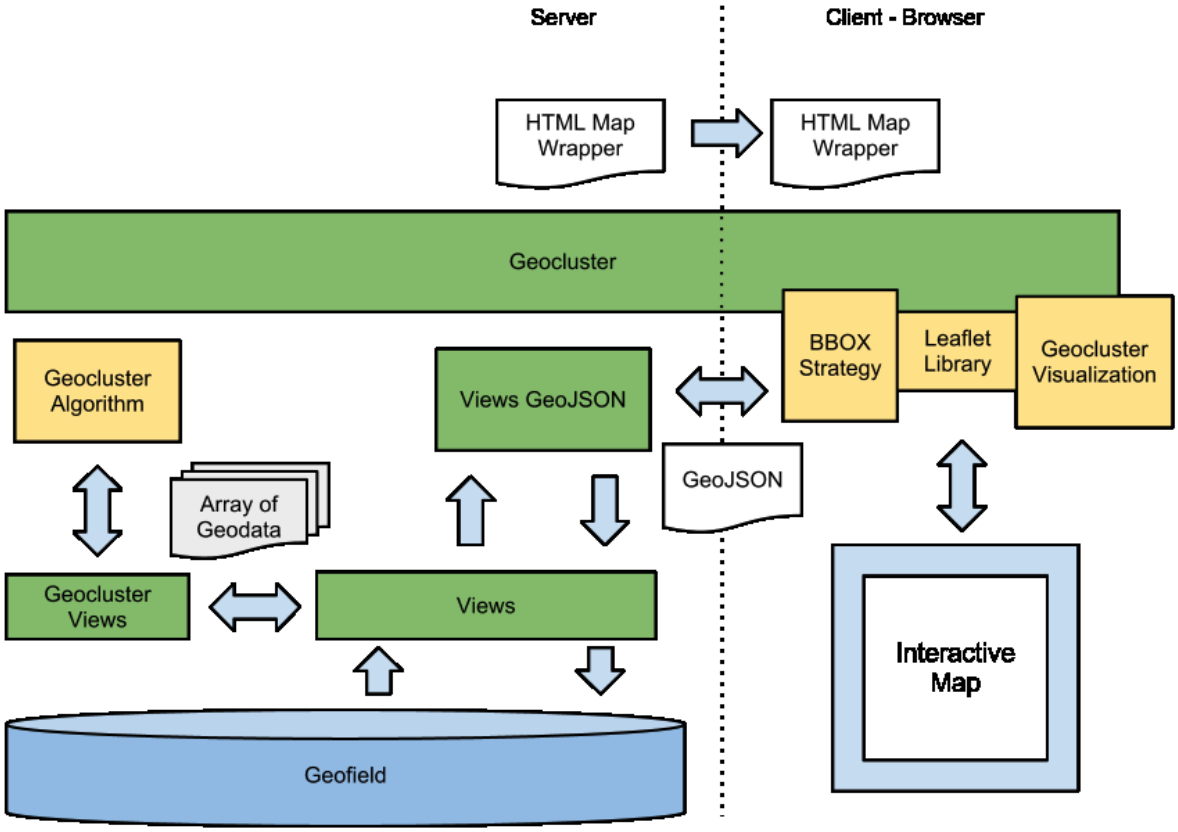
Conception & Implementation

- Algorithm
  - Geohash attributes
  - On-the-fly clustering vs. pre-calculated clusters
- Geocluster Drupal 7 module
  - Cluster geo data: Views integration
  - Deliver clustered data: Views GeoJSON integration
  - Visualize clustered data: Leaflet and optional OpenLayers integration
- SolrGeocluster Solr plugin
- Geocluster Solr integration module
- Performance benchmarks

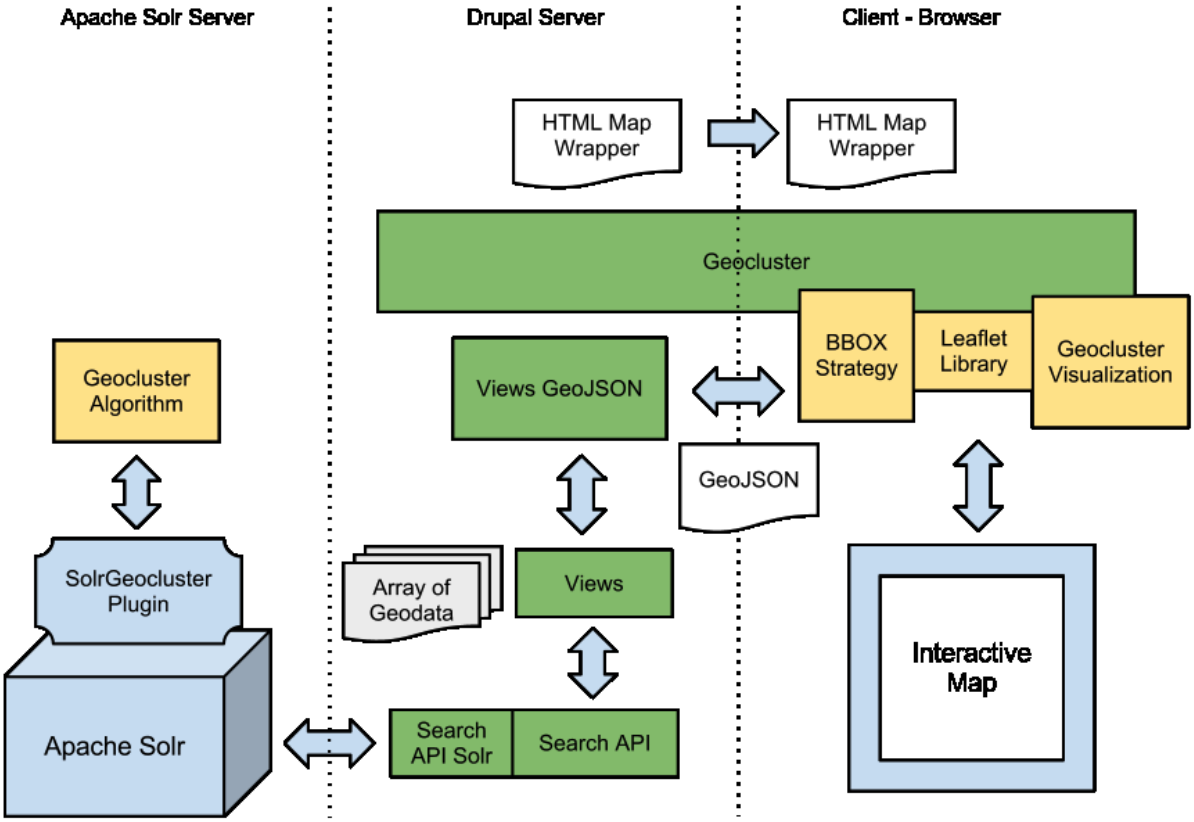
Overview Diagrams



Drupal Mapping query and display modules - Geocluster



Drupal Mapping query and display modules - Geocluster Solr



References

- Geospatial Search Using Geohash Prefixes, David Smiley, Open Source Search Conference, 2011
- Mapping with Drupal, Alan Palazzolo, Thomas Turnbull, O'Reilly Media, 2011
- Spatial clustering of structured objects, Antonio Varlaro, Università degli Studi di Bari, 2008
- W. Wang, J. Yang, and R. Muntz. ``STING: A statistical information grid approach to spatial data mining," Proceedings of the 23rd VLDB Conference, Athens, Greece, 186±195, 1997.
- Vizualizing Large Spatial Datasets in Interactive Maps, Jean-Yves Delort, Macquarie University and Capital Markets CRC, Sydney, Australia, 2010
- Hierarchical Cluster Visualization in Web Mapping Systems, Jean-Yves Delort, Macquarie University and Capital Markets CRC, Sydney, Australia
- Spatial Clustering Algorithms and Quality Assessment, Jingke Xi, School of Computer Science and Technology, 2009
- Multi-Level Clustering and its Visualization for Exploratory Spatial Analysis, Vladimir Estivill-Castro and Ickjai Lee, 2002
- Current state of technology and potential of Smart Map Browsing in web browsers, Emanuel Schütze, Bremen University of Applied Sciences, 2007
- Visualization of Time-Oriented Data By Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, Christian Tominski
- Mastering the Information Age Solving Problems with Visual Analytics Edited by Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis and Florian Mansmann
- An Efficient Web-GIS Solution based on Open Source Technologies: A Case-Study of Urban Planning and Management of the City of Zagreb, Croatia Mario MILER, Drazen ODOBASIC and Damir MEDAK, Croatia
- Hall, B., Leahy, M.G. 2008. Open Source Approaches in Spatial Data Handling, Series. : Advances in Geographic Information Science. Springer.
- Holmes, C., Doyle A., Wilson, M. 2005. Towards a Free and Open Source (FOSS) Spatial Data Infrastructure. From Pharaohs to Geoinformatics, FIG Working Week 2005 and GSDI-8, Cairo, Egypt April 16-21, 2005.
- Medak, D., Pribicevic, B., Djapo, A., Medved, I. 2003. Open Source based Spatial Data Infrastructure - Why and How? /Proceedings of the ISPRS WG VI/3 Workshop: Geoinformation for Practice, Vol. XXXIV, Part 6/W11, 193-196, Zagreb.
- Sayar, A., Pierce, M., Fox, G. 2006. Integrating AJAX Approach into GIS Visualization Web Services, Proceedings of IEEE International Conference on Internet and Web Applications and Services ICIW'06 February 23-25, 2006 Guadeloupe, French Caribbean.
- Fast Map Interaction Without Flash (MapBox) - **http://www.slideshare.net/devseed/fast-map-interaction-without-flash**