Distributed Vector based Spatial Data Conflation Services

Sérgio Freitas, Ana Paula Afonso

Department of Computer Science – University of Lisbon
Lisbon, Portugal.

sergio.freitas@novageo.com, apa@di.fc.ul.pt

Abstract. *Spatial data conflation is a key task for consolidating geographic knowledge from different data sources covering overlapping regions that were gathered using different methodologies and objectives. Nowadays this research area is becoming more challenging because of the increasing size and number of overlapping spatial data sets being produced. This paper presents an approach towards distributed vector to vector conflation, which can be applied to overlapping heterogeneous spatial data sets through the implementation of Web Processing Services (WPS). Initial results show that distributed spatial conflation can be effortlessly achieved if during the pre-processing phase disjoint clusters are created. However, if this is not possible further horizontal conflation algorithms are applied to neighbor clusters before obtaining the final data set.*

1. Introduction

The ability to combine various datasets of spatial data into a single integrated set is a fundamental issue of contemporary Geographic Information Systems (GIS). This task is known in scientific literature as spatial data conflation and is used for combining spatial knowledge from different sources in a single mean full set.

Till recent years automatic spatial data conflation research has been primarily concerned with algorithms and tools for performing conflation as single thread operations on specific types of datasets, primarily using geometry matching techniques [Saalfeld 1988] and lately semantic matching has been identified has a key element of the conflation problem [Ressler *et al*. 2009]. With the advent of Web based maps an increasing number of community and enterprise generated knowledge is being produced using heterogeneous techniques [Goodchild 2007].

The increasing size of data sets is a central aspect that spatial data conflation algorithms have to overcome and the demand to perform on the fly operations in an Internet environment. To overcome these constraints it is fundamental that conflation operations can be distributed between several computing instances (nodes) in order to complete fusion operations in satisfactory time for very large data sets.

The overall spatial conflation process is composed by five main sub-processes, analysis and comparison, preprocessing, matching, fusion and post-processing [Wiemann and Bernard 2010]. Analysis and comparison evaluates if each data set is a candidate for conflation and if further preprocessing is needed to make each data set compatible (e.g. coordinate system conversion, map alignment, generalization); after this task the matching process is used to find similar features, a combination of

geometrical, topological, semantic similarity measurements are used to find similar features and afterwards fusion is performed between candidate features; finally post-processing is applied to perform final adjustments.

A fundamental aspect for implementing geographic services is the use of Open Geospatial Consortium (OGC) standards that will allow existing GIS software packages that implement these standards to easily interact with the services being implemented.

MapReduce is a programming model developed by Google that is widely adopted for processing large data sets on computer clusters [Dean and Ghemawat 2004]. MapReduce is composed by the Map and Reduce steps. Map is responsible to sub-divide the problem and distribute to worker nodes, and then worker nodes process the smaller data set and return the results to the master node. Reduce is responsible to collect the results and combine them according to a predefined process.

In order to achieve distributed conflation, spatial clustering algorithms are applied in the preprocessing phase to each input data set so each output cluster can be matched and fused in autonomous nodes (Map). At last results from each computing instance are merged in post-processing phase in order to reach the desired final output (Reduce).

Spatial conflation service prototypes are currently being developed through the implementation of Web Processing Services (WPS) standard defined by the OGC [OGC 2007]. Apache Hadoop MapReduce framework is invoked by the WPS engine (PyWPS) to perform distributed and scalable spatial conflation. The base software components are all open source projects (PyWPS, GDAL/OGR, GEOS and PostgreSQL/PostGIS). This is a key aspect of this work because the usage of open source solutions allows the full control of each task performed and a greater knowledge of the inner works of each software component. Our initial results show that distributed spatial conflation can be easily achieved if during the preprocessing phase disjoint clusters are created ensuring that throughout the post-processing phase there is no need to apply horizontal conflation algorithms (e.g. edge-matching) to merge features that are placed on the edge of each cluster. If this is not possible further horizontal conflation algorithms have to be applied during the Reduce step before obtaining the final data set.

This paper presents an approach towards distributed vector to vector conflation, which can be applied to overlapping heterogeneous vector spatial data sets. The conflation methodologies are geared towards detecting data clusters that can be computed in independent nodes and subsequently merged.

2. Related Work

Spatial data conflation is a specialized task within geoinformatics that is mainly used for detection change, integration, enrichment of spatial data sets and updating [Yuan and Tao 1999]. Conflation is commonly classified as Horizontal or Vertical [MacMaster 1986]. Horizontal conflation is used to define conflation applied to adjacent spatial data sets, and vertical conflation is concerned with overlapping data sets [Beard and Christman 1986].

A comprehensive mathematical context for automated conflation process was firstly proposed by Saalfeld [Saalfeld 1988]. This initial work was focus on performing

feature geometries alignment between data sets. The first step of this process is to recognize similar geometries, check if matching is correct using quality control points, and apply feature geometry alignment using geometric interpolation and space partitioning algorithms. This process is applied recursively until no similar geometries were found on each data set. The main conclusion of Saalfeld's work is that Delaunay triangulation is the best fit for partitioning space and these partitioning arrangements certify that independent liner transformations (e.g. scaling and rotation) could be performed to geometries in order to align data sets inside each triangle.

This technique is described in the conflation literature as *rubber-sheeting* and is still widely used for performing alignment operations between data sets using control points that can be automatically calculate by matching features between data sets or using humans to determine common control points on each data set [White 1981].

Conflation can be applied to raster and vector data sets, and can be categorized as raster to raster, raster to vector and vector to vector conflation. Each category uses different algorithms and techniques. Raster conflation implies the use of image analysis techniques [White 1981], raster to vector involves image analysis and comparison with feature geometries, and vector to vector is focused on the analysis of geometry and feature attributes [Saalfeld 1988].

Current conflation process is composed of several sub-tasks [Wiemann and Bernard 2010]. Firstly, input data sets have to be analyzed and compared to ensure fitness for further processing tasks. This includes analyzing metadata or inferring geometrical, topological and semantic properties. Data gathered during the previous step is feeded to the pre-processing task which determines if further map alignment, coordinate system conversion or generalization has to be performed. After this task feature matching is computed using a wide range of techniques that compute geometric, topologic and semantic feature similarity. This is an important task in the conflation process. If this step is not able to achieve unambiguous mapping the whole process can be compromised or in some systems, humans are used to disambiguate uncertainty. Afterwards the fusion task is responsible for merging matched features, which includes full or partial merging of the geometric and attributes. Finally post processing is performed to attain the final output data set.

Feature matching has evolved through the years. Initially, the main focus was geometric and topology similarity [Saalfeld 1988] using simple geometrical metrics as distance, length, angle or linearity [McMaster 1986]. Afterwards attribute based feature matching was proposed using a role based approach [Cobb *et al*. 1998]. Lately feature matching has evolved to measure semantics similarity [Giunchiglia el al. 2007] based on attributes, geo-ontologies or data structures [Janowicz *et al.* 2011].

The usage of distributed spatial conflation services was proposed by [Wieman and Bernand 2009] using the WPS standard. However, these authors did not describe the distribution methodology and they only briefly refer that the use of Web Services is advantageous in the implementation of spatial conflation.

3. Conceptual Design of Distributed Conflation Services

A central aspect for successfully designing conflation services is the service ability to access spatial data from different data sources [Yuan and Tao 1999]. It is very difficult to fully support read and write operations on proprietary data formats, non-standard application programming interfaces (API), and heterogeneous metadata definitions [McKee 2004]. Even if the conflations service is able to read a subset of input data formats, other issues like acquisition methods, data structures and diverse semantic definitions can become very challenging.

To overcome these difficulties a fundamental aspect for designing conflation services is implementing OGC standards that will allow existing GIS software packages that support these standards to easily interact with the services being developed.

The WPS standard is the most suitable OGC service standard to implement conflation services. It provides rules for standardizing inputs and outputs, methods for measuring service performance and can be used to build service chains using service orchestration techniques [Wiemann and Bernard 2010]. Input data sets required by the WPS service can be delivered across a network or available at the server side [OGC 2007].

The distributed data conflation services being developed are composed by several processing services that can be chained together to complete a full conflation service (Figure 1a). The first activity that is performed is the analysis and comparison of the given input datasets in order to determine if the data sets are compatible for conflation and if further preprocessing is needed. During the preprocessing activity inconsistencies between data sets are removed by performing several tasks (e.g. map alignment, coordinate transformation, generalization) according to the requirements identified during the analysis and comparison phase. Another key task performed is the division of the input data sets in subsets that will allow the distribution of the matching and fusion activities (Figure 1b). During the matching phase similar features that represent the same object are identified in both data sets. Afterwards, it is performed the fusion of matched features. Finally, during the post-processing phase, if overlapping features area is founded on adjacent data sets, subsets are merged and horizontal conflation is performed.

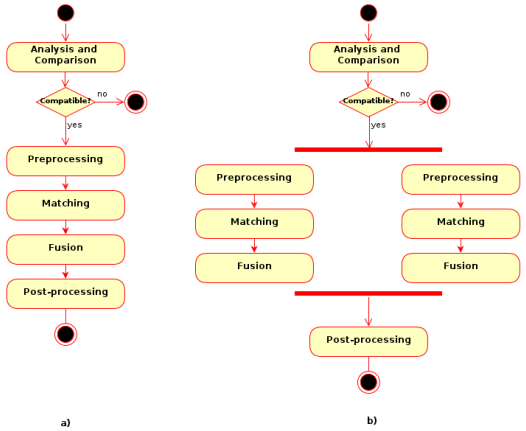

Figure 1. Conflation Services Activity Diagram

To perform distributed processing on spatial data sets the processing service has to be able to divide these data sets in subsets. Generally in distributed processing of geospatial data, tiling techniques are applied to obtain subsets that can be processed in a distributed system [Kruger and Kolbe 2008]. These techniques are based on the creation of a regular grid that divides the space according to a given measure on each dimension of the coordinate system. After obtaining the grid a simple matching algorithm is applied between the grid and the data set features to obtain all the features that are contained on each cell of the grid (Figure 2). Then, these features are considered a subset.

Using a regular grid imposes that similar features can be assigned to different grid cells. Even if input data is used to generate the regular grid, it is very difficult to obtain a grid where similar features are more likely to be maintained in the same cell.

The main difficulty of using a grid to create subsets appears when similar features are assigned to different cells, and in this case during the distributed matching phase they will not be identified and consequently fusion of these features will fail.

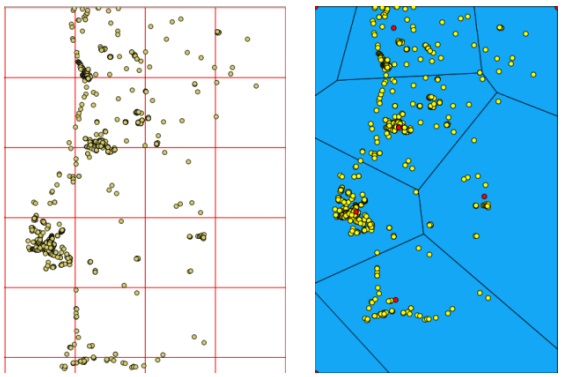

Figure 2. Tiling versus Clustering applied to OpenStreetMaps Points of Interest

To overcome this problem during the preprocessing phase clustering algorithms are applied to the input datasets in order to ensure that similar features are in the same subset. Given the increasing size of input data set only fast non fuzzy clustering algorithms are being considered, namely Web-Scale K-Means Clustering [Sculley 2010] and DBScan [Ester *et al.* 1996]. After applying these clustering algorithms to input datasets a Voronoi Tessellation [Franz 1991] is performed to define the shapes that will be used to extract each subset (Figure 2).

4. Implementation

To build a proof of concept we are using open source based software. This is an important aspect of this work because the usage of open source solutions allows the full control of each task and a greater knowledge of the inner works of each software component.

The development of the WPS service is being performed using the PyWPS project, a widely used Python based WPS engine. All spatial data processing algorithms are based on OGR and GEOS libraries. Data storage is performed using

PostgreSQL/PostGIS, and Apache Hadoop MapReduce framework is invoked by the WPS engine to perform distributed and scalable spatial conflation.

Distributed conflation services deployment is performed on the Amazon Web Services (AWS) cloud based environment. The ability to create new computing instances on demand is used to create nodes to perform Map/Reduce operations on the Hadoop MapReduce framework.

A simple distributed point conflation service was developed using the software stack described above. This first service implementation uses fast k-means for data clustering, Euclidean distance for measuring geographic similarity and string based attribute comparison for attribute matching. Features fusion is achieved using the average between each similar feature spatial position and a full merge of feature attributes.

This service will be further developed to support lines and polygons using clustering algorithms adapted to this type of features and different distance calculations techniques.

5. Conclusions

The developed concept and the simple implementation of point conflation service has demonstrated that distributed vector based conflation services are feasible and the use of clustering algorithms to create subsets can improve the performance of the feature matching and fusion process on a distributed conflation service.

The definitions of the WPS service interface are important to achieve a greater abstraction and independence between the service being developed and the clients. This allows a greater interoperability because changing the underlying development and deployment methods does not affect service usage.

Initial results show that distributed spatial conflation can be effortlessly achieved if during the pre-processing phase disjoint clusters are created. However, if this is not possible further horizontal conflation algorithms are applied to neighbor clusters before obtaining the final data set.

The developed distributed conflation services will be used to evaluate if the presented approach is better fitted to perform distributed conflations than using gridding techniques to create subsets.

Current research is focused on reaching a base conflation service design that can be used to perform distributed conflation on a cloud based environment. After this initial phase each service activity will be further developed to increase the overall conflation performance.

References

Kruger, A. Kolbe, T. (2008). “Mapping spatial data infrastructures to a grid environment for optimized processing of large amounts of spatial data”, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXXVII, Beijing, China.

Cobb, M. Chung, M. Miller, V. Foley, H. Petry F., and Shaw K (1998). "A Rule-Based Approach for the Conflation of Attribute Vector Data", GeoInformatica, 2(1), 7-35.

Dean, J. and Ghemawat, S. (2004) "MapReduce: Simplified Data Processing on Large Clusters", In: 6th Symposium on Operating Systems Design and Implementation, San Francisco, USA.

Ester M. Kriegel, H. S, J. and Xu X. (1996) "A density-based algorithm for discovering clusters in large spatial databases with noise". In: Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining.

Franz A. (1991) "Voronoi diagrams – A Survey of a Fundamental Geometric Data Structure". In: ACM Computing Surveys 23(3), 345-405.

Giunchiglia, F. Yatskevich, M. and Shvaiko P. (2007) "Semantic Matching: Algorithms and Implementation" In: Journal on Data Semantics IX, Springer-Verlag, Berlin, 1-39.

Goodchild M. (2007) "Citizen and sensors: the world of volunteered geography". GeoJournal 69, p. 211-221. Springer Science+Bussiness Media.

Janowicz K., Raubal M. and Kuhn W. (2011) "The Semantics of Similarity in Geographic Information Retrieval", In: Journal of Spatial Information Science, 2, 29-57.

McKee, L. (2004) "The Spatial Web", White Paper, Open GIS Consortium.

McMaster, R. (1986) "A Statistical Analysis of Mathematical Measures for Linear Simplification", In: The America Cartographer, 13, 103-116.

OGC (2007). "OpenGIS Web Processing Services". Open Geospatial Consortium Editions, Version 1.0.0.

Ressler J., Freese E. and Boaten V. (2009) "Semantic Method of Conflation". In: Terra Cognita 2009 Workshop In Conjunction with the 8th International Semantic Web Conference. Washington, USA.

Wiemann S., Bernard L. (2010) "Conflation Services within Spatial Data Infrastructures". In: 13th Agile International Conference on Geographic Information Science 2010. Guimarães, Portugal.

White, M. (1981). The Theory of Geographical Data Conflation. Internal Census Bureau draft document.

Saalfield, A. (1998) "Conflation: Automated Map Compilation". International Journal of Geographic Information Systems, 2(3), 217-228.

Sculley, D. (2010) "Web-scale K-Means Clustering". In: Proceedings of WWW 2010.