A Prototype System for Value-Range Queries

Ruixin Yang, Kwang-Su Yang, Jiang Tang, Menas Kafatos
Center for Earth Observing and Space Research (CEOSR)
School of Computational Sciences
George Mason University
Fairfax, VA 22030
ryang@gmu.edu

Abstract-We have developed algorithms for value-range queries of Earth science data with an area size constraint. The value-range condition is approximately satisfied by a histogram clustering technique with a data pyramid model. The result is the high resolution areas (points) on which the value range condition is satisfied. The area size condition is achieved by combining convex hull and point-in-polygon algorithms and depth-first strategy to reconstruct areas. Those areas are constructed with areas selected by the value-range queries.

Using a data pyramid concept, the above two algorithms can be considered as drill-down searching and bottom-up reconstruction procedures. This paper describes a prototype system based on above algorithms. The web-based prototype system shows the procedures as the queries being answered. On the server side, we are using Splus software with consideration of programs in general languages such as C and FORTRAN and DBMS for efficiency. On the client side, SVG (Scalable Vector Graphics) and Java techniques are merged to have a clear and user-friendly interface.

I. INTRODUCTION

The Earth observing from space, in particular, NASA's Earth Observing System (EOS), computational weather and climate models, and other Earth observations are producing massive data product at a rate of TB/day level [1]. Efficient methods are needed for knowledge extractions and even for searching useful data from such huge volume of data. To extract knowledge from Earth science data, users must first investigate the data content. The so-called content-based queries are also important for users to select interesting data before ordering from a data center with vast data holdings.

One simple content-based query is to find areas over which the parameter values fall in a given range. The term, value-range query, may be more appropriate in this case and is used with content-based query without distinction. The query result could be used for ordering data as well as for defining features associated with scientific concepts. A data pyramid model was proposed for efficiently answering this kind of content-based queries [2]. Yang *et al*. [3] limited the pyramid level to two and associated the cells on the lower level with histograms and histogram based clusters. The query is answered approximately with a given and guaranteed accuracy.

Another more complex content-based query is similar to the above simple query but with more constraints. The constraints will make the query result more specific for certain scientific usages. For example, Yang *et al*. [4] and Yang *et al*. [5] considered value-range queries with conditions on area size and proportion of missing value points. The convex hull and point-in-polygon algorithms are used in this extension.

To make the content-based query more widely accessible, we developed a web-based prototype for using the algorithms developed already. In the following section, we summarize the details about the two kinds of value-range queries. Then, in section III, we describe the prototype system with examples. In section IV, we discuss the result and potential future work.

II. VALUE RANGE QUERY ALGORITHM

Suppose one has a griddded data set which contains values of a geophysical parameter over a certain area for a specific time or time period. A simple value range query is to find spatial areas (and/or temporal ranges) over which the parameter values fall in a given range. For example, in NASA' s CIDC data sets [6], there is a NDVI data set which gives the monthly global 1^ox1^o NDVI values. A simple value range query is to find all the 1^ox1^o areas on which the monthly NDVI values are between, say 0.04 and 0.26, for a specific month.

Li *et al*. [2] built data pyramids bottom up by using statistic values such as minimum, maximum, mean, etc. When a value-range query is given, one will drill down the data pyramid to search the lowest resolution data first, and then search again in the next level on the selected areas and so on until the user-specific resolution or the pyramid bottom is reached. In this case, the value-range queries are answered approximately with hope that certain percentages of such areas are found.

Yang *et al.* [3] improved this method by limiting the pyramid to two levels only with guaranteed accuracy. The original given data are treated as the higher resolution data. The lower resolution cells are created based on the higher resolution areas, and each cell is associated with one histogram built on the higher resolution data values over that cell. One characteristic of Earth science data is that the spatial distribution is not uniform. Actually, similar values are found in the near neighbor areas but significantly different values often exist in distant areas. This property is used to cluster histograms based on value distributions. Therefore, cells on which the value distributions are similar are grouped together and indexed accordingly. Furthermore, histograms in the same group are summed up to form a representative histogram for that cluster.

When a value-range query needs to be answered, the system will compute the contribution from each cluster by checking the representative histograms and sort all the contributions. Usually, after picking just a few clusters that contribute the most to the given value range, we may include most of the searched areas. More precisely, we can select more than the given percentage, say 95%, of higher resolution areas from the cells associated with just a few groups. To obtain the higher resolution areas, we can scan the higher resolution areas in the selected lower resolution cells only. By doing this, we only need to work on a much reduced numbers of cells, therefore, speed up the process.

Again, we use the NDVI example. We divide the cells into six groups by a clustering algorithm. We want to find 95% areas in the higher resolution data over which NDVI values are between 0.04 and 0.26. We compute the contribution based on the representative histograms and find that the three clusters that contribute the most would include 41%, 32%, and 25% such areas, respectively. Therefore, we guarantee more than 95% the queried areas would be included by scanning the higher resolution areas in cells in the three groups only. The result is shown in Figure 1 and Figure 2, respectively. Figure 1 gives the lower resolution cells selected (green). In Figure 2, the higher resolution areas in the selected lower resolution cells are displayed by the dark green areas.

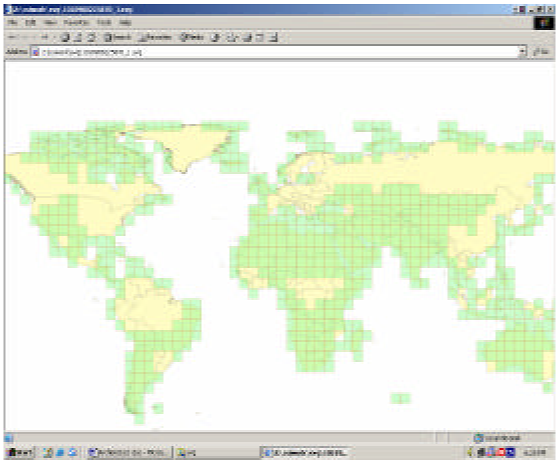


Figure 1. Selected lower resolution cells (green).

From Figure 2, we can conclude that most selected higher resolution areas are spatially contiguous. We believe that most contribution from isolated areas is filtered out by the 95% requirement. Even though, there exist a few isolated small areas selected by above algorithm. When a scientist use the value-range query result to define features, it is more likely the scientist will consider the areas of large sizes instead of the isolated small areas. Therefore, it is beneficial to separate the two kinds of areas by imposing other constraints to the simple value-range queries.

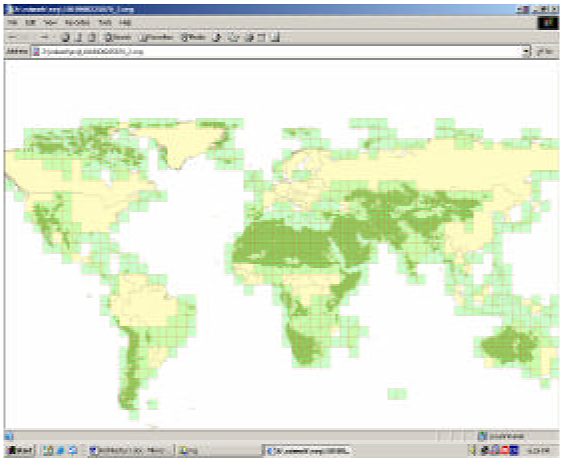


Figure 2. Selected higher resolution areas (dark green).

The value-range queries with area size constrains are studied by the same group [4, 5]. In addition to the value-range condition described above, we also ask the result satisfying the area size and missing data proportion conditions. For example, consider the NDVI data again, we want to find areas with the following conditions: the average NDVI value is between 0.04 and 0.26; the missing value proportion is less than 15%; and the valid data points are not less than 100 (mimicking the size condition).

Convex hull and point-in-polygon algorithms [7, 8] are integrated together to develop a reliable, feasible, depth-first method for answering the value-range queries with area size constraints. The result for the NDVI example is shown in Figure 3. The areas satisfying all conditions are the shaded polygons. There are overlaps between some shaded polygons, which is a problem to be overcome in the future. It is obvious that small isolated areas over which the NDVI value ranges fall in the given range are not selected due to the violations of the area size conditions.

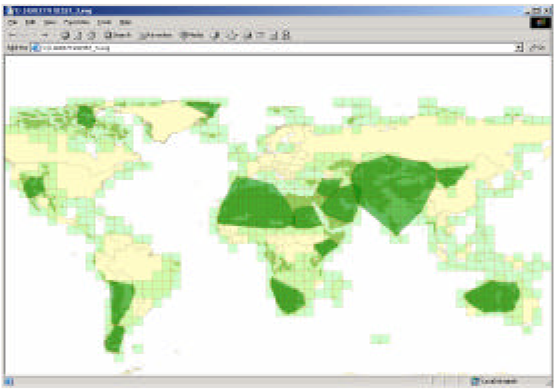


Figure 3. Convex hulls satisfying the area size conditions (shaded).

III. PROROTYPE SYSTEM

We develop a web-based prototype system, which integrates above algorithms with the web technology. Figure 4 shows a potential system architecture for a distributed data information system supporting the value-range queries. In this design, we utilize new technologies such as Geography Markup Language (GML) [9] and the Scalable Vector Graphics (SVG) [10] to leverage existing and future software components. At present, we only developed the following components, the server for handling data in file systems, the SVG converter that transforms the server output directly into SVG without using GML, and a Java-based client. Those available components are denoted by the green color in the architecture diagram.

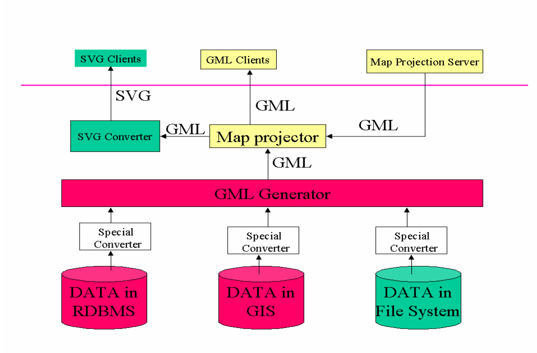
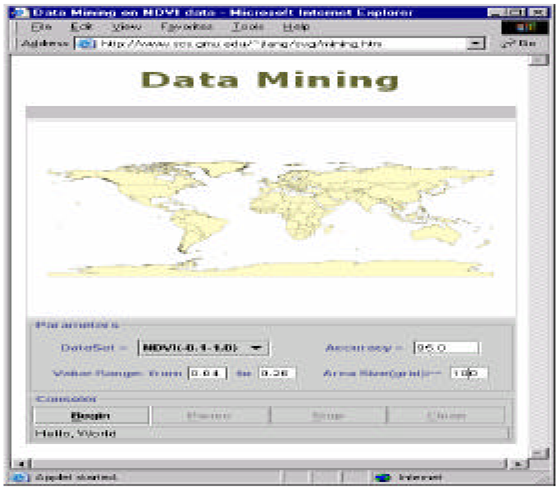


Figure 4. Potential prototype architecture.

On the server side, a simple version of Map Projector was implemented and it supports a linear scalar projection as default as well as the Plate-Carree Projection. This Map Projector converts the retrieved data from searching engine to the designated projection, which later is used by the SVG converter to generate an SVG-formatted file.

The SVG, a language from W3C “… *for describing two-dimensional vector and mixed vector/raster graphics in XML* …” [10], is adopted here because it is a standard vector graphics format for web browsers. At this moment, on the client side, an SVG viewer plug-in must be installed since SVG is not supported as a default format in current web browsers. Moreover, a Java applet is developed as another client component to control the query processes and communications among different system components. Through the Java/SVG interface, users can set query parameter values, launch a retrieval session in the server, and display the result (SVG file) phase by phase, leaving to users the ability to start, pause, resume, and terminate the query session as well.

Figure 5 displays the Java applet with the continent boundaries in SVG as the graphics background. A user, via the applet, can select a parameter, say NDVI, input the value range, say, 0.04 and 0.26, give the minimum accuracy, say 95%, and limit the minimum area size, say 100 data points in this special case. After the user sends the query to the server, the server will answer the query and send back the result step by step. The first result is the lower resolution cells selected as those in Figure 1. Then, the higher resolution areas in the selected cells are checked and those satisfying the value range condition are sent back to the client and displayed as shown in Figure 2. Finally, the result for the query with area size conditions is returned as given in Figure 3.


Data Mining on NDVI data - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address
Go
Data Mining
Parameters
Console
Begin
Pause
Stop
Clear
Applet started.
Internet

DataSet = NDVI(-0.1-1.0)

Accuracy = 95.0

Value Range: from 0.04 to 0.26

Area Size(grid)= 100

Hello, World

Figure 5. User interface.

During the process, the user can interrupt, continue or terminate the session on his/her wish for studying the details. This session control mechanism will be very useful when the full query needs a relatively long time to answer, and users may be satisfied by the middle result already and does not want to wait for the final outcome. In addition, on the client side, it is very easy to zoom in/out the images because it is a built-in function in SVG plug-in.

IV. DISCUSSION

The prototype system for the value-range query and the SVG-based client shows a promising technology combination. The SVG allows users to easily study the result in their own ways, and the interactive control applet provides a mechanism for users to decide what to see and what not to see.

The weakness of the current prototype is that the response time on the server side is slow. The slowness is due to the complexity of the algorithms as well as the software utilized. We are searching and studying programs written in C or FORTRAN for the same tasks and are working on integrating the low level programs with the prototype system. A significant improvement on the system response speed is expected by using such programs.

Another interesting work is to use DBMS such as Oracle to support above algorithms and the prototype. Oracle has built-in procedures for histograms and Spatial Data Option for handling spatial data. It seems that Oracle cannot support the clustering process directly although it supports histogram creations and histogram-based queries. More research is needed to effectively and efficiently use Oracle for such an integrated task.

ACKNOWLEDGEMENT

This work is partially supported by NASA ESTO (NAG5-8607 and NAG5-10177).

REFERENCES

[1] G. Asrar and R. Greenstone, eds., 1999 "EOS Reference Handbook," NASA (Washington, D.C.), 1999.

[2] Z. Li, X. Wang, M. Kafatos, R. Yang, 1998, "A Pyramid Data Model for Supporting Content-based Browsing and Knowledge Discovery", Proceedings of the Tenth International Conference on Scientific and Statistical Database Management, pp. 170-179. IEEE, Computer Society.

[3] R. Yang, K. Yang, M. Kafatos, X. Wang, 2000. "Value Range Queries on Earth Science Data via Histogram Clustering," Proceedings of International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining (TSDM2000), September, 2000

[4] Kwang-Su Yang, Ruixin Yang, Menas Kafatos, "A Feasible Method to Find Areas with Constraints Using Hierarchical Depth-First Clustering", in Proceedings of the 13th International Conference on Scientific and Statistical Database Management (L. Kerschberg and M. Kafatos, eds.), pages 257-262, IEEE, Computer Society, 2001.

[5] Ruixin Yang, Kwang-Su Yang, Menas Kafatos, "A Clustering Technique for Value-Range Queries with Area Size Constraints," In Proceedings of ESTO meeting, Paper A9P3, 2001

[6] Kyle, H.L., McManus, J.M., Ahmad S., et al., 1998., "Climatology Interdisciplinary Data Collection, Volumes 1-4, Monthly Means for Climate Studies," NASA Goddard DAAC Science Series, Earth Science Enterprise, National Aeronautics & Space Administration, NP-1998(06)-029-GSFC.

[7] J. O'Rourke, "Computational Geometry in C," Cambridge University Press, 1994.

[8] F. P. Preparata, M. I. Shamos, "Computational Geometry," Springer-Verlag, 1985.

[9] Ron Lake, Adrian Cuthbert, eds, "Geography Markup Language (GML) v1.0," http://www.opengis.org/techno/specs/00-029/GML.html, 2000.

[10] Jon Ferraiolo, eds, "Scalable Vector Graphics (SVG) 1.0 Specification," http://www.w3.org/TR/SVG/index.html, 2001