A High Performance Web-Based System for Analyzing and Visualizing Spatiotemporal Data for Climate Studies

Zhenlong Li, Chaowei Yang, Min Sun, Jing Li, Chen Xu,
Qunying Huang, and Kai Liu

Center of Intelligent Spatial Computing for Water/Energy Sciences, George Mason University, Fairfax, VA, 22030-4444
{zli1,cyang3,msun,jlih,cxu3,qhuang1,kliu4}@gmu.edu

Abstract. Large amount of data are produced at different spatiotemporal scales by many sensors observing Earth and model simulations. Although advancements of contemporary technologies provide better solutions to access the spatiotemporal data, it is still a big challenge for researchers to easily extract information and knowledge from the data due to the data complexities of high dimensions, heterogeneity, distribution, large amount and frequently updating. This is especially true in climate studies, because climate data with coverage of the entire Earth and a long time period (such as 200 years) are often required to extract useful climate change information and patterns. A well-developed online visual analytical system has the potential to provide an efficient mechanism to bridge this gap. Using performance improving techniques for an online visual analytical system, we researched and developed a high performance Web-based system for spatiotemporal data visual analytics includes the following components: 1) a Spatial Data Registration Center for managing the big spatiotemporal data and enabling researchers to focus on analyses without worrying about data related issues such as format, management and storage; 2) a workflow for pre-generating and caching frequently requested data to reduce the server response time; and 3) a technique of "single data fetch, multiple analyses" to reduce both server response time and client response time; Finally, we demonstrate the effectiveness of the prototype through a few use cases.

Keywords: Big Data, CyberGIS, Online Visual Analytics, Computing Optimization.

1 Introduction

Over the past decades, the advancements of sensor technologies have greatly improved our capabilities to record the spatiotemporal snapshots of a variety of natural or social phenomena as data. The spatiotemporal data collected by these sensors are characterized by high dimensions, heterogeneity, distribution and large amount (Yang et al. 2010). In climate studies, scientists use the spatiotemporal data captured by various sensors to validate and improve climate models. Meanwhile, large amounts of (or big) spatiotemporal data are generated by climate model runs. For example,

S. Liang, X. Wang, and C. Claramunt (Eds.): W2GIS 2013, LNCS 7820, pp. 190–198, 2013.

ModelE (Schmidt et al., 2006), one of the climate models, can simulate the climate of Earth system in many scenarios, and will produce approximately 25 gigabytes data with a single run for a 30-year monthly simulation. Generally, climate studies require hundreds, thousands, to millions of model runs with simulations of 200 years, and will generate petabytes of spatiotemporal data. It is an urgent scientific demand to effectively manage and mine the big spatiotemporal data to support scientific research.

With the advancements of the Internet and distributed computing technologies, many visual analytic applications are migrated from standalone computing environment to Web-based computing environment. There are several initiatives of online visual analytic system in climate domain. For example, the TRMM (Tropical Rainfall Measuring Mission) online Visualization and Analysis System (TOVAS) is developed by NASA (Berrick et al. 2009) to analyze TRMM gridded rainfall data. Fetch Climate application (http://fetchclimate.cloudapp.net/) is developed by Microsoft to access mean values of different climate parameters for selected geographic regions of the Earth surface.

Online visual analytical systems make it more convenient for end users in that they do not have to install the entire analytical software package on local computers, nor to download and prepare their own data locally. However, performance becomes an essential issue when developing online visual analytic systems (Yang et al. 2011). An acceptable waiting time for a Web system response is approximately 3-8 seconds (Corner 2010). In a simple Web application, this second rule is easy to comply. However, when manipulating big spatiotemporal data, system performance should be considered systematically similar to the WebGIS performance (Yang et al. 2005).

This paper addresses Web-based system performance in two aspects: 1) how fast the system is? This measures the system response time between sending the requests to receiving the results by end users, 2) how interactive and intuitive the system is? This is hard to quantify but can be evaluated by the user interface complexity (Coskun et al. 2005), which could be reduced by minimizing the number of user clicks and reducing the number of Web pages involved in an operation (Galitz 2007). Focusing on these two performance aspects, we developed a high performance online visual analytical system.

2 Related Work

Online visual analytics of big spatiotemporal data provides a potential solution and is becoming increasingly important (Liu et al. 2007). For example, Sun et al (2012) developed a Web-based visualization platform for climate studies based on Google Earth. Their approach provides a state of the art visualization style, but the installation of Google Earth plug-in significantly reduces the portability and availability. Another limitation for this approach is that it does not support user-defined areas (or Area of Interest, AOI), which poses a major issue when mining knowledge and patterns at a local or regional scale. Liu et al (2009) developed an online analysis system focusing on global satellite precipitation algorithm validation and inter-comparison. While

these initiatives provide good examples for developing online visual analytical applications in climate domain, the critical system performance challenge is barely mentioned considering massive concurrent requests, complex processing, and large amounts of datasets (Yang et al. 2011).

Metadata are descriptive data about the data and have been widely used to manage large amounts of data in data-intensive applications (Yee et al. 2003, Singh et al. 2003, Weibel et al. 1998). A spatiotemporal visual analytical system is a typical data-intensive scientific application for managing and analyzing Terabytes to a Petabyte of data that may span millions of files or data objects. Singh et al (2003) proposed a Metadata catalog service to support general data-intensive applications. We adopt this approach in the SDRC by extending the spatiotemporal support to improve data discovery performance.

Caching data on both the server side and the client side can reduce the load on network transmission and server processing (Rowstron et al. 2001, Li et al. 2011, Yang et al. 2005) and improve the overall system performance. There are two types of caching: 1) caching data based on user requests on-the-fly with dynamic cache architecture and 2) loading at start-up and caching preprocessed data with static cache architecture (Yang et al. 2005). Leveraging the static cache architecture, we propose a workflow to pre-generate and cache frequently requested data to reduce the server response time.

By adopting the approaches discussed above, we developed a high performance Web-based spatiotemporal visual analytical system for climate studies. The system performance is optimized from various aspects, including server side data storage, management and preprocessing, client side data presentation, client/server data transmitting and user interactive interface. This paper elaborates the details of the system: section 3 introduces the system overall architecture; section 4 explains the methods and techniques used in the system; section 5 conducts three experiments to evaluate the performance of the proposed methods and techniques; section 6 demonstrates the system with two usage scenarios to show how it can facilitate climate studies in practice; section 7 concludes and discusses future research.

3 System Architecture

Our high performance online visual analytical system includes the spatiotemporal data registration center (SDRC), intermediate result pre-generator, server side advanced analyzer and client side analyzer (Fig.1):

SDRC is the core component for server side data management. Raw data and pre-generated data from Intermediate Result Pre-generator are registered into SDRC. Metadata is extracted and recorded in the Metadata Database dynamically whenever a new dataset is registered. Metadata and spatiotemporal index techniques are adopted in this component.

The Intermediate Result Pre-generator is responsible for pre-generating intermediate result for frequently accessed data to avoid repeated calculations and reduce server response time. Caching and data pyramid techniques (Yang et al. 2005) are adopted in this component.

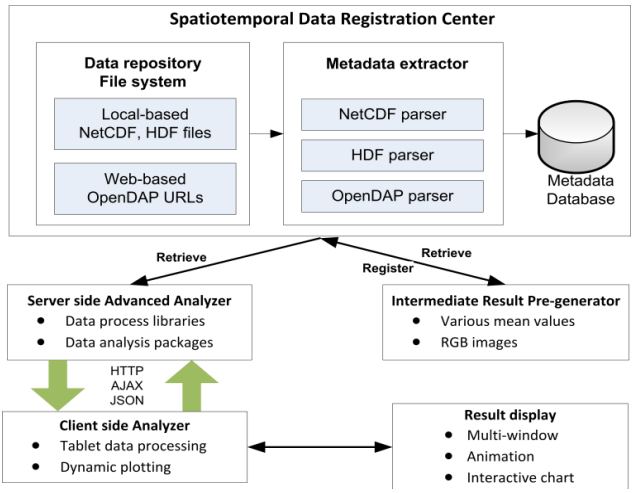

Fig. 1. System Architecture

Server Side Advanced Analyzer is a function-driven component including third party libraries and packages: NetCDF-Java (http://www.unidata.ucar.edu) java-based library and R package (http://www.r-project.org/).Client Side Analyzer is responsible for 1) providing an effective mechanism "single data fetch, multiple analyses" to reduce server response time and user interactive time, 2) processing and plotting the returned table data.

4 Components and Methods

This session elaborates the system components and methods from three aspects: server side data management, client side result presentation and server/client data communication.

4.1 Spatiotemporal Data Registration Center

The major goal of SDRC is to manage all the registered spatiotemporal data in a uniformed manner by dynamically extracting the metadata of the spatiotemporal data into a Metadata Database. Data repository maintains all the registered data in the file system, data may either be manually input to a directory, or in most cases, be uploaded from distributed model runs automatically. Metadata extractor is activated whenever a new dataset is uploaded to the data repository. The plug-and-play parsers of NetCDF (Network Common Data), HDF (Hierarchical Data Format) and OpenDAP (Open-source Project for a Network Data Access Protocol) interpret the data when adding, extract metadata and insert to the Metadata Database. To enhance the security and interoperability, a standard Web service is provided when accessing metadata outside of SDRC (Erl 2004).

The SDRC is flexible, extensible and interoperable in that 1) the metadata is automatically extracted to the database whenever new data are registered, 2) the structure

of Metadata Extractor could support any kinds of data formats by adding corresponding plug-and-play parsers, and 3) the standard Web service based on REST (Representational State Transfer) (Costello 2007) serves a secure method for accessing the metadata though the Web in an interoperable manner.

4.2 Intermediate Result Pre-generator

In climate studies, most analyses are based on averaged (mean) values of monthly, quarterly or yearly data spanning a long time period. These values are either calculated for every data point or for a point-of-interest region with averaged values of all the data points. Since the mean values are the basic inputs for various advanced analyses, it is necessary to do the calculation only one time and store them in a proper manner for further access to avoid repeating calculations. Visualizing the spatiotemporal data as a two dimensional image is an effective way for analysts or general users to visually discover spatial patterns (Keim 2000). Because these images are frequently requested by the users, pre-generating them could dramatically reduce the server response time, especially when animating for a long time period.

Therefore, we proposed a workflow used in Intermediate Result Pre-generator component (Fig.2) to 1) pre-calculate the intermediate mean values, 2) pre-render frequently accessed images, 3) effectively store and manage these pre-generated data for fast access. Three types of intermediate data are pre-generated: temporal-based means stored in NetCDF files, spatial-based means stored in CSV files, and RGB images rendered from both original data and the newly created mean data (NetCDF). Finally, all these newly generated data are registered in SDRC.

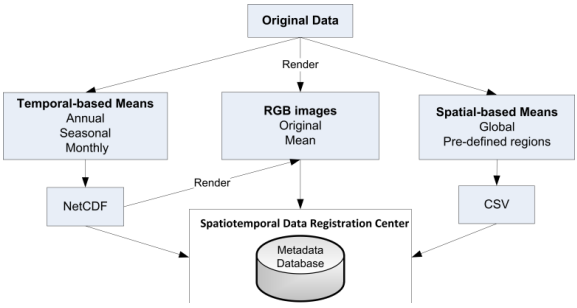

Fig. 2. Workflow for pre-generating frequently requested data

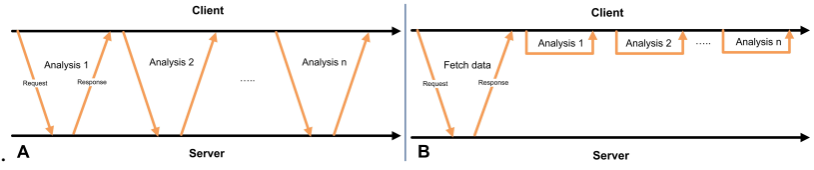

Fig. 3. (A) Traditional online analysis workflow; (B) Proposed online analysis workflow

4.3 Single Data Fetch, Multiple Analyses

A typical spatiotemporal analyses request includes five parameters: dataset(s), variable(s), time period, spatial coverage and analysis type. The analysis type includes time series, correlation, animation, anomaly and outlier to name a few. For the traditional approach (Fig.3A), a request containing the five parameters is built for each analysis. This approach is not efficient or interactive because 1) for each analysis, server side dataset (s) need to be read at least one time, which is time consuming, 2) only statistic images are returned to the client, any modification to the result, for example, adjust a parameter or to change a plot style, needs a new request-wait-response workflow.

To overcome these drawbacks, we use a "single data fetch, multiple analyses" (Fig.3B) method. This method can significantly improve the overall system performance by 1) reducing the number of request-wait-response loops by carefully considering both the request parameters and the response data structure; 2) shifting partial of the visual analytical processes from the server side to the client side. Instead of returning static images, the processed data are returned to the client side for further analysis and visualization.

5 Performance Comparison

Three experiments were conducted to evaluate the performance of the proposed components and methods.

5.1 Performance of SDRC

The data used in this experiment are generated from ModelE runs with 10 years' monthly data from 1951 to 1960. Nine NetCDF datasets with total data volume of 80 Gigabytes are generated. 566 climatic variables are included in these datasets. Figure 4A shows the server response time with and without SDRC when extracting the variables, time period and spatial coverage from the data in a Web-based environment. When using SDRC, variables, time period and spatial coverage are fetched directly from the metadata database instead of being extracted from the nine NetCDF files, which dramatically reduces the server response time.

5.2 Performance of Intermediate Result Pre-generator

This experiment requested the 1 year mean data, 5 years mean data, global mean of 5 years mean data and RGB image of 5 year mean data with or without using pre-generating/caching method. Data used is the same as the previous experiment as described in 5.1. As presented in figure 4B, rendering "1 year mean images" is the most time consuming process. Since "1 year mean" need to be calculated before rendering "1 year mean images", most time is consumed by calculating "1 year mean" (more than 20 seconds). Time consumed by "5 years mean" and "5 years global mean" is relatively shorter, but we can still see a significant performance improvement after using pre-generating/caching method.

5.3 Performance of "Single Data Fetch, Multiple Analyses"

In this experiment, we use three variables and two AOIs from one data to 1) conduct time series analysis for each variable in each study area, 2) for each study area, conduct correlation analysis for any 2 of the 3 variables. We recorded the time consumed on the server side and the number of request-wait-response loops for these analyses when using and without using "single data fetch, multiple analyses" method. The result (Tab.2) shows that "single data fetch, multiple analyses" could dramatically reduce the total server response time by reducing the number of workflow loops.

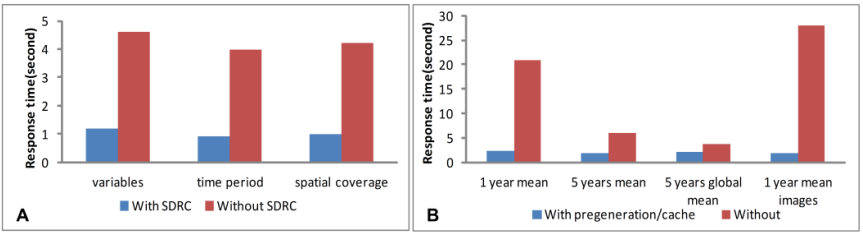

Fig. 4. (A) Server response time with and without SDRC; (B) Server response time with and without pre-generation/cache

Table 1. Request/response loops and consumed time with/without "single data fetch, multiple analyses" method

Analysis type		With	Without
Time series	Request/response loops	1	6
	Total time (second)	4.3	12.6
Correlation	Request/response loops	1	21
	Total time (second)	4.2	41.7

6 System Demonstration

For time series analyses, users can select multiple model simulations, multiple variables and multiple AOIs at the same time, and plot them together for better comparison (Fig.5A).

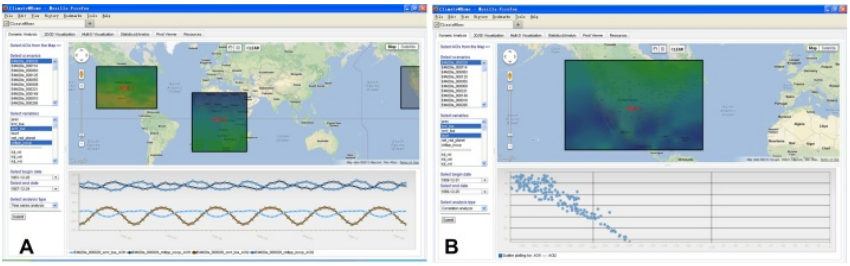

Fig. 5. (A)Time series plotting for two variables and two AOIs ; (B) Correlation analyses for two variables

Correlation analyses enable users to plot the relationships of any two variables or two AOIs. Figure 5B shows the scatter plot for two variables from one simulation at the same area of interest.

7 Discussion and Conclusion

This paper developed a high performance web-based system to analyze and visualize spatiotemporal data to support climate studies by focusing on the performance improvement methods and techniques. The system performance is optimized by leveraging performance improvement techniques of server side data management and data preprocessing, client side result presentation and user interactive interface, and server/client communication.

Three experiments were conducted using ModelE data to evaluate the performance of the proposed methods and techniques. The results show that these methods and techniques could dramatically improve the system performance by reducing the system response time and user interactive time. A prototype is developed based on the proposed methods and techniques. The prototype is served as an initial approach to handle big spatiotemporal data in climate domain.

Further studies are required to continue improving the system through, but not limited, to the following four aspects: 1) improving SDRC by leveraging the techniques used in big data management (Herodotou et al.2011, Bughin et al.2010); 2) enhancing the spatiotemporal data storage and process strategies by leveraging spatial cloud computing (Yang et al. 2011); 3) optimizing the system architecture for collecting, archiving, sharing, analyzing and visualizing spatiotemporal data by adopting geospatial cyberinfrastructures technologies to support various scientific domains (Zhang et al. 2009, Yang et al. 2010); 4) providing more innovative result display methods based on the characteristics of spatiotemporal data and scientific requirements.

Acknowledgements. Research reported is supported by Microsoft Research, NSF (IIP-1160979), and NASA (NNX12AF89G).

References

1. Bailey, B.P., Konstan, J.A., Carlis, J.V.: The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In: Proceedings of INTERACT, vol. 1, pp. 593–601 (2001)
2. Berrick, S., Leptoukh, G., Liu, Z., Pham, L., Rui, H., Shen, S., Teng, W., Zhu, T.: Multisensor distributive on-line processing, visualization and analysis system. In: Proceedings of the 2004 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2004, vol. 3, pp. 2030–2033 (2004)
3. Bughin, J., Chui, M., Manyika, J.: Clouds, big data, and smart assets: Ten tech-enabled business trends to watch. McKinsey Quarterly 56 (2010)
4. Corner, S.: The 8-second rule, `http://www.submitcorner.com/Guide/Bandwidth/001.shtml`
5. Coskun, E., Grabowski, M.: Impacts of User Interface Complexity on User Acceptance and Performance in Safety-Critical Systems. Journal of Homeland Security and Emergency Management 2(1) (2005)

6. Erl, T.: Service-oriented architecture: a field guide to integrating XML and web services. Prentice Hall PTR (2004)
7. Hendler, J.: Web 3.0 Emerging. Computer 42(1), 111–113 (2009)
8. Herodotou, H., Lim, H., Luo, G., Borisov, N., Dong, L., Cetin, F.B., Babu, S.: Starfish: A self-tuning system for big data analytics. In: Proc. of the Fifth CIDR Conf. (2011)
9. Keim, D.A.: Designing pixel-oriented visualization techniques: Theory and applications. IEEE Transactions on Visualization and Computer Graphics 6(1), 59–78 (2000)
10. Li, Z., Yang, C.P., Wu, H., Li, W., Miao, L.: An optimized framework for seamlessly integrating OGC Web Services to support geospatial sciences. International Journal of Geographical Information Science 25(4), 595–613 (2011)
11. Liu, Z., Rui, H., Teng, W., Chiu, L., Leptoukh, G., Kempler, S.: Developing an online information system prototype for global satellite precipitation algorithm validation and intercomparison. Journal of Applied Meteorology and Climatology 48(12), 2581–2589 (2009)
12. Liu, Z., Rui, H., Teng, W., Chiu, L., Leptoukh, G., Vicente, G.: Online visualization and analysis: A new avenue to use satellite data for weather, climate, and interdisciplinary research and applications. In: Measuring Precipitation From Space, pp. 549–558 (2007)
13. Rowstron, A., Druschel, P.: Storage management and caching in PAST, a large-scale, persistent peer-to-peer storage utility. ACM SIGOPS Operating Systems Review 35(5), 188–201 (2001)
14. Schmidt, G.A., Ruedy, R., Hansen, J.E., Aleinov, I., Bell, N., Bauer, M., Yao, M.S.: Present-day atmospheric simulations using GISS ModelE: Comparison to in situ, satellite, and reanalysis data. Journal of Climate 19(2), 153–192 (2006)
15. Singh, G., Shishir, B., Ann, C., Ewa, D., Carl, K., Mary, M., Sonal, P., Laura, P.: A metadata catalog service for data intensive applications. In: 2003 ACM/IEEE Conference on Supercomputing, p. 33. IEEE (2003)
16. Sun, X., Shen, S., Leptoukh, G.G., Wang, P., Di, L., Lu, M.: Development of a Web-based visualization platform for climate research using Google Earth. Computers & Geosciences 47, 160–168 (2012)
17. Galitz, W.O.: The essential guide to user interface design: an introduction to GUI design principles and techniques. Wiley (2007)
18. Taylor, K.E.: Summarizing multiple aspects of model performance in single diagram. Journal of Geophysical Research 106(7), 7183–7192 (2001)
19. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery. In: Internet Engineering Task Force RFC 2413, vol. 222 (1998)
20. Yang, C., Goodchild, M., Huang, Q., Nebert, D., Raskin, R., Xu, Y., Bambacs, M., Fay, D.: Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing? International Journal of Digital Earth 4(4), 305–329 (2011)
21. Yang, C., Wu, H., Huang, Q., Li, Z., Li, J.: Using spatial principles to optimize distributed computing for enabling the physical science discoveries. Proceedings of the National Academy of Sciences 108(14), 5498–5503 (2011)
22. Yang, C., Wu, H., Huang, Q., Li, Z., Li, J., Li, W., Sun, M., Miao, L.: WebGIS performance issues and solutions. In: Advances in Web-Based GIS, Mapping Services and Applications. Taylor & Francis Group, London (2011) ISBN 978-0
23. Yang, C., Raskin, R., Goodchild, M., Gahegan, M.: Geospatial cyberinfrastructure: past, present and future. Computers, Environment and Urban Systems 34(4), 264–277 (2010)
24. Zhang, T., Tsou, M.H.: Developing a grid-enabled spatial Web portal for Internet GIServices and geospatial cyberinfrastructure. International Journal of Geographical Information Science 23(5), 605–630 (2009)