Global Workload Characterization of A Large Scale Satellite Image Distribution System

Brian Romoser [1], Ribel Fares [2], Peter Janovics [3], Xiaojun Ruan [4], Xiao Qin [5], Ziliang Zong [6*]

[1,2,3,6]*Computer Science Department, Texas State University*
[4]*Department of Computer Science, West Chester University of Pennsylvania*
[5]*Department of Computer Science and Software Engineering, Auburn University*
[1,2,3,6]{bromoser, rf1190, pwj3, zz11}@txstate.edu
[4]xruan@wcupa.edu [5]xqin@auburn.edu

***Abstract*—Online content distribution systems, which store incredibly large amounts of information and provide service to large numbers of users, are becoming increasingly commonplace. To fulfill the wide range of requests sent by different users, these systems must ensure efficient handling of massive amount of data. To achieve this goal, the in-depth analysis and comprehensive understanding of user behaviors are critical. However, analyzing the behaviors of worldwide users with different needs is a very challenging task. This is especially true when historical user behaviors evolve over time or may be affected by unpredictable events. In this paper, we present a number of workload characterization techniques applied to one of the world's largest online satellite image distribution systems operated by the U.S. Geological Survey (USGS) and NASA.**

***Keywords*-clustering, content distribution system, geospatial data, geovisualization**

I. Introduction

The Information Age is so named after the ever-increasing amounts of data available on a daily basis [1]. Consequently, with an expanding data set comes a heightened need for the efficient management of information flow. Among the myriad of powerful new consumer software fueled by this data explosion are geospatial applications such as Google Earth [2] and Microsoft MapPoint [3]. These programs rely on large stores of geographic information to provide a wide range of utilities to their users, such as Google Earth's ability to provide detailed satellite images of nearly any location on Earth. Without efficient workload management systems, these new applications would be unable to handle the large amount of operations they generate.

In April 2008, the United States Geological Survey (USGS) Earth Resources Observation and Science (EROS) center opened their online satellite imagery archive freely to the public after successfully serving researchers privately for many years. The Global Visualization Viewer (GloVis) [4] enables users to browse and download terrestrial images captured by the National Aeronautics and Space Administration (NASA) Landsat program from 1972 to the present [5], with data provided by the system put to use in fields as diverse as agricultural development, regional management, education, and many government applications. Increased demand for images places high importance on quick preparation and distribution of content to users, but the system's processing time of up to days for non-cached images elevates efficiency to paramount importance. The performance of conventional caching algorithms on the EROS system has been discussed in previous publications[6], but significant performance gains have yet to be achieved. Other methods such as employing modified caching strategies and utilizing generic data mining techniques have been proved unable to provide sizable performance improvements.

For such a large-scale system, efficiency improvements would yield substantial savings in terms of time and power usage. A good understanding of the workload characteristics plays an important role in deciding which caching strategies and other performance improvement techniques to adopt. The workload generated by GloVis constantly changes in several dimensions as new users make requests in response to real-world events and 250+ new sets of image data are downlinked daily, providing a continuous feed of new content to be requested.

In this paper, we present an in-depth analysis of user, image, and request characteristics of the USGS EROS system. We present the results of the characterization of over five million lines of real-world data and use the information therein to reveal insightful patterns and glean useful knowledge on user behavior.

We make use of Google Earth's abilities to display user generated data to assist our discovery of meaningful relationships within our data set. The usefulness of Google Earth as a geovisualization aid has been explored previously [6], [7], [8], but there do not yet exist many studies featuring its use in assisting in workload characterization.

The remainder of this paper is organized as follows. Section II contains background information on USGS EROS and the GloVis system, as well as the motivation for characterizing the workload of the GloVis servers. Section III briefly reviews the conventional methods that have been applied to the EROS data set previously. We present the results of our experiments in section IV and discuss the implications of our findings. Within section V, we summarize our findings and explain the significance thereof, as well as lay the foundation for future work related to the EROS GloVis system.

978-1-4673-4883-6/12/$31.00 ©2012 IEEE

II. Background

In this section, we will provide further information on the EROS system that is the subject of this paper, including its unique environmental constraints. We will detail the log files provided by EROS and the manner in which they were processed to provide the material for our research. Finally, we will review the important terms used when describing the geospatial data captured with the Landsat satellites.

A. EROS Log File

The USGS-provided log file which we received contained over 5,000,000 lines of data, with each line representing one user request for one image to be processed and made available for download. Figure 1 contains a snapshot of the log file received from EROS, showcasing the raw data it contained. From each request, we extract the ID of the user generating the request, the date on which the request was made, and an agglomerate variable which represents the location on the globe and the acquisition date of the image being requested. A further description of the log file components may be seen in Figure 2.

B. Worldwide Reference System

Image data warehoused by EROS is cataloged using the Worldwide Reference System (WRS), a global notation system designed for referencing Landsat data. The two satellites still in operation, Landsat 5 and Landsat 7, use an extension of the WRS system known as WRS-2. There exist technical distinctions between the two reference systems in order to account for the large orbital differences between the older and most recent launches in the Landsat missions, making WRS and WRS-2 coordinates incompatible without conversion [5]. Despite the differences between the WRS protocols, both systems attempt to divide the globe into a grid mirroring the satellites' orbital paths, with each area on the grid referred to as a scene.

1) Scene: WRS-2 features a global grid that divides Landsat 5 and Landsat 7's orbital tracks into a system of 248 latitudinal segments known as **rows** (North-South) and 233 longitudinal segments called **paths** (East-West). The satellites' irregular orbits are responsible for a small amount of overlap across neighboring scenes, and only terrestrial imagery is captured, bringing to total number of unique scenes available to approximately 17,000. Each scene is targeted for imaging by the satellites once every orbit (approximately every 16 days), providing a third dimension of time. A unique combination of row, path, and acquisition date is referred to as an image.

2) Image: Landsat 5 and 7 both complete 14.5625 orbits per day, and as such, complete one 'viewing' of the earth every 16 days. Thus, any given scene should receive new imaging downlinked approximately every 8 days. The frequency at which images are acquired means that every scene gains about 47 new images every year. Prospective users of the EROS data must request one or more scenes using the respective WRS-2 row/path combination (forming the X and Y axes) with the addition of an acquisition date, exemplified in Figure 3.

III. Conventional Methods

In general, traditional caching strategies such as LRU and LFU yield major improvements compared to FIFO [9], [10], [11], [12], [13]. However, previous work shows that conventional caching algorithms can only reach approximately 45% of hit ratio on the EROS system [6]. We implemented a market basket analysis (MBA) based prefetching scheme to further improve the system. Images that are often requested together were processed together. The improvements were very low (less than .01%) for both LRU and LFU.

In a massive global system like EROS, it is likely that many different types of user behavior patterns are present. Efficient processing of different patterns may require different strategies; there may be no single approach suitable for all such patterns. Characterizing the workload based on suitability-of-strategy could shed light in the creation of more complex caching and/or prefetching algorithms.

IV. Characterization

Gaining a more complete view of the data set supplied by USGS provided the driving force behind our employ of workload characterization techniques. We seek the ability to classify portions of the imagery available through EROS. Through intelligent classification, images may be automatically selected as likely candidates for future requests by a prefetching algorithm with the intent of avoiding expensive cache misses upon a real request coming from a user. When we begin to characterize the workload described in the logs provided by EROS, we discover the existence of multiple trends in data requests that could be used for classification.

By exploring these unique relationships, new knowledge may be gathered on both user behavior and system performance. In this section, we will present three dominant trends in the log data we chose to monitor throughout the course of our analysis. The perspectives covered will be user-oriented, scene-oriented, and request-oriented. When examining the server logs from a user-oriented perspective, a small group of users is found to make tens of thousands of image requests, while most users only make a few. In addition, when taking a scene-centric perspective, there exist few scenes which may be considered highly popular whereas the large majority of scenes are not. Scenes with a cumulative download request total exceeding 1000 will now be referred to as *popular* scenes. Finally, in exploring the request perspective, it is seen major historical events such as natural disasters have a significant impact on users' behavior. We shall discuss the possibility of exploiting these three workload characteristics using caching and prefetching strategies.

A. Characterizing Users

USGS/EROS users may be characterized as aggressive and casual users [6]. Within the system trace, the aggressive behavior of few users persists. Figure 5 shows that top 100 aggressive users account for almost the same amount of requests made by the remaining 79,437 casual users. As the green line in Figure 4 illustrates, aggressive users create a

```
"INVENTORY_ID","CONTACT_ID","PROD_CODE","PROD_DESCRIPTION","DATE_ENTERED"
"LE70750891999325EDC01",455762,"D210","L7 ETM+ REFLECTIVE BROWSE DOWNLOAD",01-APR-12 00.00
"LE70760901999332EDC00",455762,"D210","L7 ETM+ REFLECTIVE BROWSE DOWNLOAD",01-APR-12 00.00
"LE70720891999272SGS00",455762,"D210","L7 ETM+ REFLECTIVE BROWSE DOWNLOAD",01-APR-12 00.00
"LE70730911999247EDC01",455762,"D210","L7 ETM+ REFLECTIVE BROWSE DOWNLOAD",01-APR-12 00.00
"LE70160372012092EDC00",337858,"D201","L7 ETM+ L1T/L1GT/L1G SLC-OFF DOWNLOAD",01-APR-12 00.00
"LE70160382012092EDC00",337858,"D201","L7 ETM+ L1T/L1GT/L1G SLC-OFF DOWNLOAD",01-APR-12 00.00
"LE70160402012092EDC00",337858,"D201","L7 ETM+ L1T/L1GT/L1G SLC-OFF DOWNLOAD",01-APR-12 00.00
```

Fig. 1. Snapshot of EROS-provided log file.

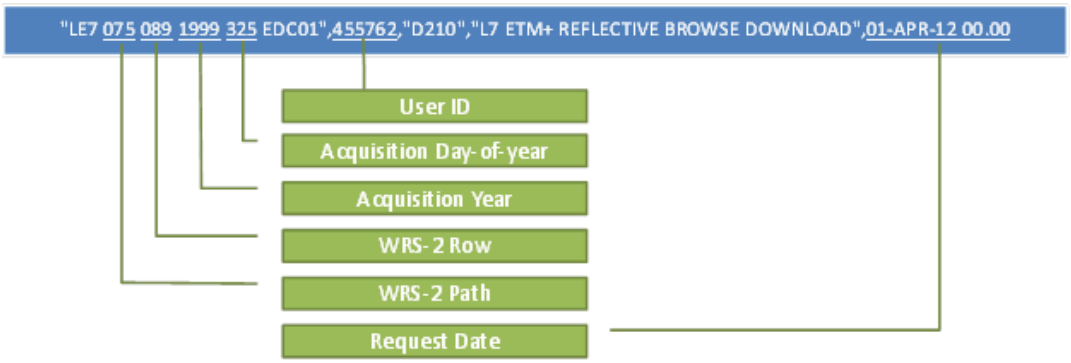

Fig. 2. Explanation of log file contents.

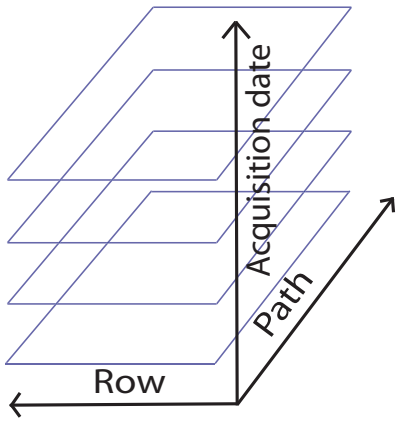

Fig. 3. Scenes are comprised of layers of images for a given row and path with respect to time.

sequential pattern in download requests for a scene as they request a wide range of imagery for a single scene. The use of sequential prefetching techniques may be able to target users that exhibit such aggressive behavior and mitigate the impact they have on the image processing queue.

B. Characterizing Scenes

While some scenes are in high demand, many other scenes are only requested a few times. Figure 6 shows that the distinction between *popular* and unpopular scenes. We also look for the same behavior at the image level. Figure 7 shows that a similar popularity distribution exists for images.

Some images and scenes are significantly more *popular* than others, but the set of *popular* scenes changes over time. In Figures 8 and 9, we compare the top 100 list of *popular* scenes and images before and after January 1st, 2011, which is about the midpoint for our data. Between the shifting time periods, 51 of the top 100 scenes remained in the list over time. On the other hand, only 8 images remained in top 100 in the same two time. This continually shifting popularity suggests that window-based caching algorithms may be better suited for the effective handling of image requests.

Findings suggest that users are more interested in newer images of the same scenes. Are there *current popular* scenes that users request the newest images available as soon as they become available? Are there *archival popular* scenes that users are mostly interested in older images? Such scenes that attract frequent download requests are referred to as *current*. Prefetching new images as they become available may be beneficial in the case of *current* scenes, however, an LFU-like approach could be more suitable for *archival* scenes.

When image requests are examined as in Figure 4, a distinct slope is seen on the rightmost points in the graph as the request date increases (moves forward in time). The tightly clustered nature of the leading edge of the graph in Figure 4 implies that requests are frequently submitted for the most recent image available at any given time for a scene. Improving upon our definition of *current* images, we consider requests for an image that are made within 60 days from the date the image is first made available as *current* images. Figure 10 captures the number of unique recent image requests for each *popular* scene. We can clearly observe clusters of scenes that have substantially more requests for *current* images than most

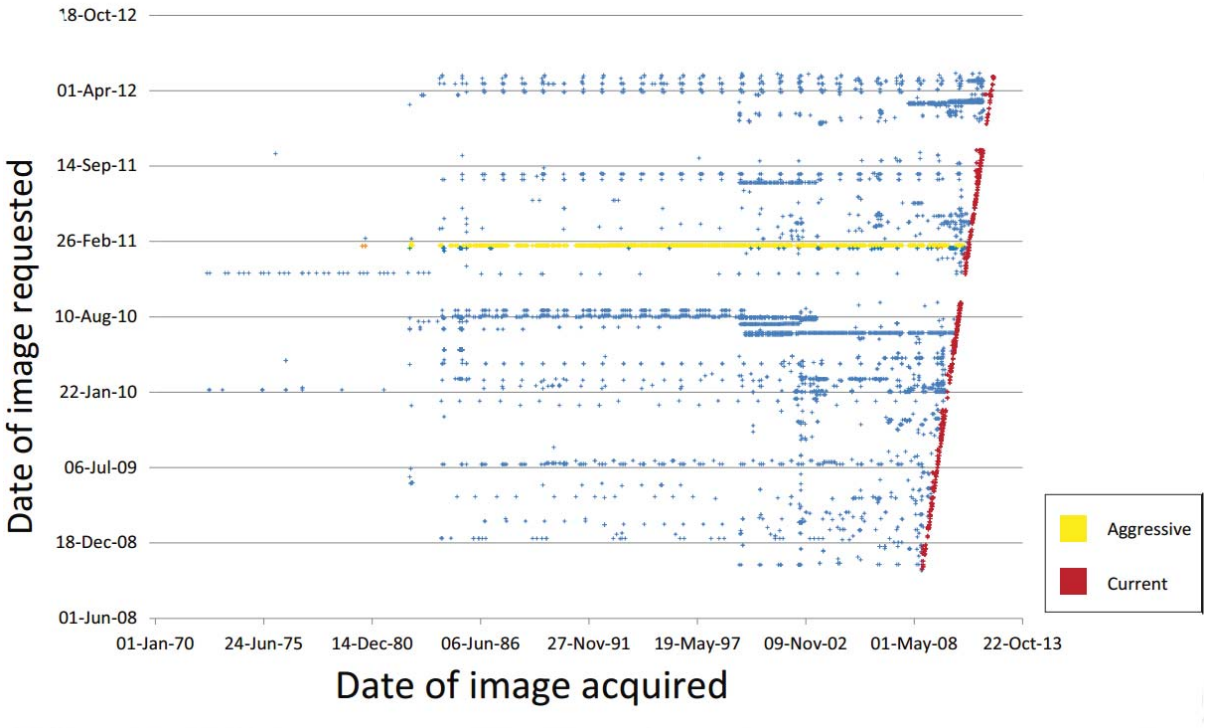

Fig. 4. Scatter plot for a coastal scene in California.

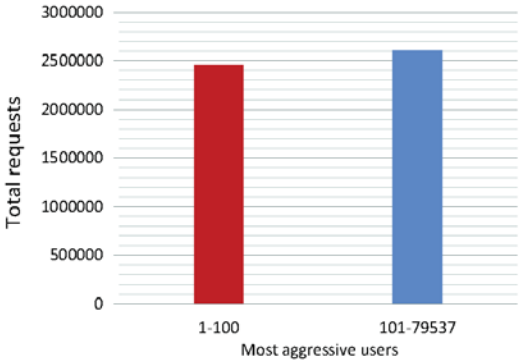

Fig. 5. Impact of top 100 aggressive users.

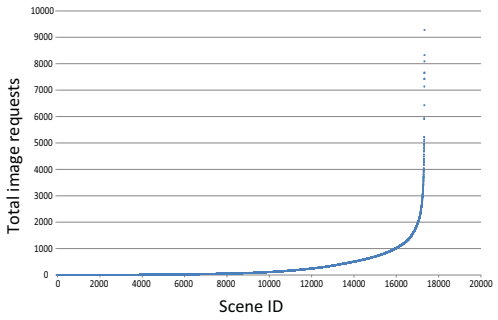

Fig. 6. Popularity distribution of **scenes.**

other scenes.

We utilize Google Earth's geovisualization abilities to examine *current* scenes. When we begin marking scenes on the map that fit the criteria for being *current*, we note that the top eight most heavily requested *current* scenes cluster along the Southwestern coast of the United States, as shown in Figure 11. As we continue to visualize the most *current popular* scenes, we find that the majority of scenes fall within the United States. We can see how *current* scenes cluster into the rough shape of the United States in Figure 12.

Figure 11 provides another valuable insight into trends in scene requests in that some scenes exhibit behavior with the

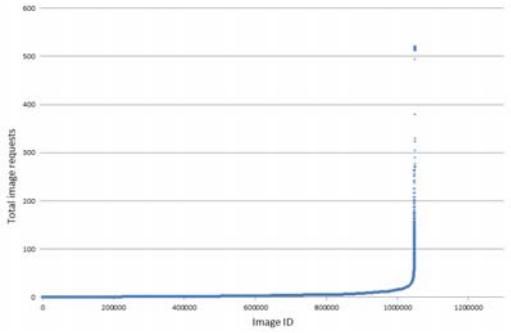

Fig. 7. Popularity distribution of **images.**

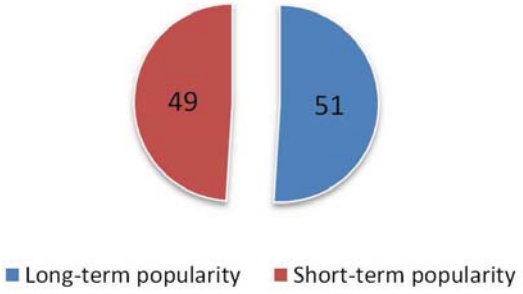

Fig. 8. Scenes stay *popular* over time.

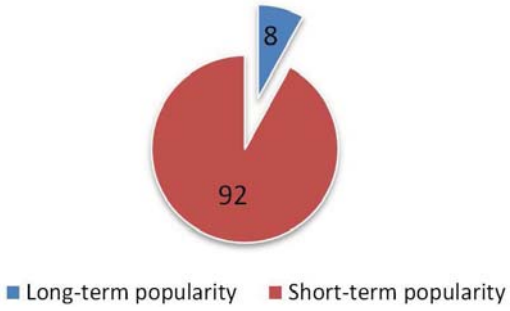

Fig. 9. Images lose popularity over time.

Fig. 10. Distribution of *current* and *archival popular* scenes.

Fig. 11. Top eight *current popular* scenes

Fig. 12. *Current popular* images with over 60 unique *current* requests.

opposite characteristics of *current* scenes - recently captured images are very rarely requested. While the majority of scenes such as northwest of Brazil in Figure 13 have a moderate number of requests for recent images, some scenes such as Brazil contain almost no such requests. Figure 13 lacks a visible slope as the x-axis is traversed, implying that the scene in question would benefit much less from any attempt to improve access to the most *current* imagery it contains. We refer to scenes that contain few requests for recent imagery yet have a large number of overall requests (200 or more) as *archival*.

When visualizing characteristics of *archival* scenes, we note several clusters appear within geographical boundaries wherein bordering scenes all share the property of being *archival*, as observable in Figures 14 and 15 below. As *current* scenes would benefit from expedited preperation of recently taken images, so too would *archival* scenes would realize performance gains by improving access to more historical imagery.

C. Characterizing Requests

Images are sometimes requested unexpectedly, and we observe multiple instances in which a particular row and path exhibit an extremely sharp peak in request density without any previous indications of doing so. Examining the scenes that express such behavior reveals that real-world events occur within the scene immediately prior to the increase in interest.

Recalling that EROS is a system utilized primarily by researchers, we find that large-scale geophysical events garner a sizable increase in requests for scenes in which they occur. The epicenter of the 2010 Haiti earthquake, located at 18°27' N, 72°31' W (row 47, path 9) [14], proved to be a scene of great interest, as seen in Figures 16 and 17. While there were zero requests for the scene containing Port-au-Prince prior to the earthquake on 12 January, 2010, an immediate surge of requests can be observed in the wake of the event, lasting many days.

Similarly to the 2010 Haitian event, the 2011 Tohoku earthquake in Eastern Japan with an epicenter located at 38°19' N, 142°22' E (row 33, path 106) [15] generated a marked increase in interest in the scene containing Tohoku,

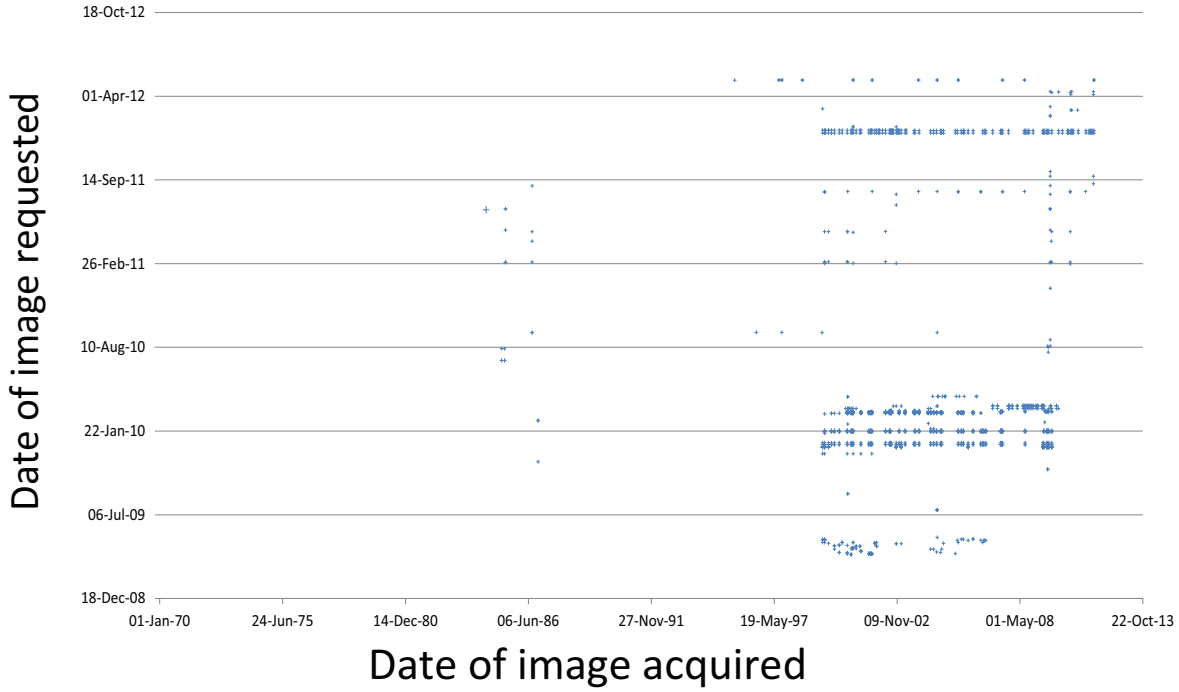

Fig. 13. Scatter plot for a scene northwest of Brazil.

Fig. 14. *Archival* popular scenes with more than 200 cumulative requests and less than 2 unique *current* requests.

Fig. 15. *Archival* popular images with more than 1200 cumulative requests and less than 5 unique *current* requests.

as well as neighboring scenes, visualized in Figures 18 and 19. Real-world events are a powerful force for motivating user download requests, even of scenes which would not have been considered *popular* before the event's occurrence. Maintaining an increased awareness of ongoing events worldwide would allow the EROS system to anticipate scenes which are likely to be the recipients of sudden bursts of requests. A prefetching system that employs event-awareness could reduce the delays caused by request processing and image preparation affecting these conditionally *popular* scenes.

Furthermore, we take note of the characteristics of requests for some scenes where major events had taken place in the past. The scene at row 24, path 26 contains the city of Pripyat, Ukraine, as well at the site of the Chernobyl Nuclear Power Plant at 51°23' N, 30°5' E [16]. The Chernobyl plant experienced a large explosion on 26 April, 1986, which is one of only two of the most highly rated events on the International Nuclear Event Scale. Although the GloVis system was not in operation at the time of the Chernobyl disaster, NASA's Landsat satellites were in orbit, capturing and archiving the imagery for future use.

Figure 20 contains requests for imagery of Chernobyl plotted to explore the relationship between the date the image was captured and the date the image was requested. There

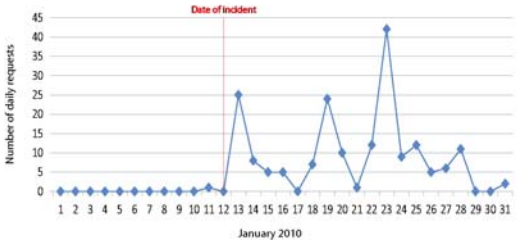

Fig. 16. Requests for Haiti in January 2010.

Fig. 17. Location of Haiti earthquake - January 2010.

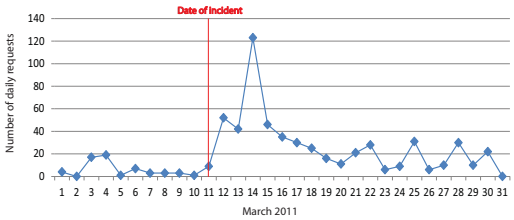

Fig. 18. Requests for Japan in March 2011.

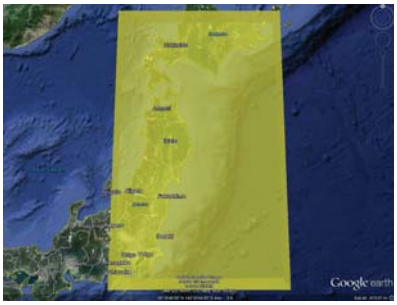

Fig. 19. Location of Japan tsunami - March 2011.

exists a vertical cluster slightly to the right of the line denoting the date of the Chernobyl disaster, signifying that the image taken on 29 April, 1986 is requested many times across the 3+ years that the GloVis server logs cover. Because this image maintains a high level of interest even as time passes, it could benefit overall system performance to reserve an amount of space in cache for this and other scenes that contain such historically significant areas of interest.

V. Conclusion and Future Work

In this paper, we have provided an explanation as to why traditional caching strategies do not work on a massive global system. Using suitable visualization techniques, we were able to identify distinct characteristics of the USGS EROS global satellite image distribution system workload.

- Few users request many images while many users only request a few.
- Some scenes are very popular while most of them are unpopular.
- For current popular scenes, users are interested in the newest available images. For archival scenes, users are mostly interested in older images.
- Hierarchical clustering of current and archival scenes reveals meaningful geographical shapes.
- Very few images are very popular.
- The popularity of scenes slowly evolves over time while the popularity of images evolves rather quickly.
- Some image requests are triggered right after important global events.
- Images of extreme historical events stay popular after many years.

Such overlapping patterns should be identified and targeted accordingly. In particular, evolving patterns should be targeted using time window-based caching strategies. An appropriate window size should be determined based on how quickly the pattern is evolving. Popularity is shown to be a bad measure for prefetching new images as they become available. Current popularity seems to be a better criterion for prefetching.

EROS currently utilizes an array of hard disk drives to serve as a cache for processed images awaiting download. By employing an intelligent prefetching policy targeting high-priority images, current levels of performance may be achievable with a significantly smaller overall cache size. By reducing the number of active hard drives needing to be kept in constant operation, EROS could realize a reduction in net power usage and ecological impact as a result of daily operations.

Acknowledgment

The authors sincerely appreciate the comments and feedback from the anonymous reviewers. Their valuable discussions and thoughts have tremendously helped in improving the quality of this paper. The work reported in this paper is supported by the U.S. National Science Foundation under Grants No. CNS-0915762, CNS-1212535, CNS-0917137, and the Texas State University Library Research Grant. We also gratefully acknowledge the support from the U.S. Geological

Fig. 20. Scatter plot for the scene encompassing the epicenter of the Chernobyl disaster outside of Prypiat, Russia.

Survey (USGS) Earth Resources Observation and Science (EROS) Center.

REFERENCES

[1] P. Lyman, and H.R. Varian, "How Much Information", 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 in May, 2012.

[2] http://www.google.com/earth/.

[3] http://www.microsoft.com/mappoint/.

[4] http://glovis.usgs.gov/.

[5] Landsat Project Science Office, "Landsat 7 Science Data Users Handbook".

[6] R. Fares, B. Romoser, Z. L. Zong, M. Nijim, and X. Qin, "Performance Evaluation of Traditional Caching Policies on A Large System with Petabytes of Data", *in Proceedings of the 7th IEEE International Conference on Networking, Architecture, and Storage*, 2012.

[7] Z. L. Zong, J. Job, X. Zhang, M. Nijim, and X. Qin, "A Case Study of Visualizing Global User Download Patterns Using Google Earth and NASA World Wind", *Journal of Applied Remote Sensing*, 2012.

[8] G. D. Standart, K.R. Stulken, X. Zhang, and Z. L. Zong, "Geospatial Visualization of Global Satellite Images with Vis-EROS" *Environmental Modeling & Software*, vol. 26, pp. 980-982, 2011.

[9] H. Chou and D. Dewitt, "An Evaluation of Buffer Management Strategies for Relational Database Systems", *in Proceedings of the International Conference on Very Large Databases (VLDB)*, 1985.

[10] S. Dar, M. Franklin, B. Jonsson, D. Srivastava, and M. Tan, "Semantic Data Caching and Replacement", *in Proceedings of the International Conference on Very Large Databases (VLDB)*, 1996.

[11] N. Megiddo and D. Modha, "ARC: A Self-Tuning, Low Overhead Replacement Cache", *in Proceedings of the USENIX Conference on File and Storage Technologies (FAST)*, pp. 115-130, 2003.

[12] Y. Zhou, J. Philbin, and K. Li, "The Multi-Queue Replacement Algorithm for Second Level Buffer Caches", *in Proceedings of the USENIX Technical Conference*, 2001.

[13] D. Willick, D. Eager, and R. Bunt, "Disk Cache Replacement Policies for Network File Servers", *in Proceedings of the International Conference on Distributed Computing Systems (ICDCS)*, 1993.

[14] http://toolserver.org/~geohack/geohack.php?pagename=2010_Haiti_earthquake&params=18.457_N_72.533_W_

[15] http://toolserver.org/~geohack/geohack.php?pagename=2011_T%C5%8Dhoku_earthquake_and_tsunami&params=38.322_N_142.369_E_

[16] http://toolserver.org/~geohack/geohack.php?pagename=Chernobyl_disaster&params=51_23_23_N_30_05_57_E_