Target-Independent Mining For Scientific Data: Capturing Transients and Trends for Phenomena Mining

Thomas H. Hinke, John Rushing, Heggere Ranganath and Sara J. Graves

Information Technology & Systems Laboratory and Computer Science Department,
University of Alabama in Huntsville,
Huntsville, AL 35899 U.S.A.
thinke@cs.uah.edu, jrushing@cs.uah.edu , ranganat@cs.uah.edu, sgraves@cs.uah.edu

Abstract

This paper describes a data mining approach for extracting enriched data from scientific data archives such as NASA's Earth Observing System Data and Information System (EOSDIS) that are stored on slow access tertiary storage. This enriched data has significantly smaller volume than the original data, yet preserves sufficient properties of this data such that over time, many different users can repeatedly mine it for different Earth-science phenomena. This enriched data captures daily trends and significant deviation from trends for each bin of gridded data from an equal-degree grid covering the Earth's surface. A feature of this enriched data is that it is independent of any particular target phenomena, although it assumes that such phenomena are either transient in nature or characterized by trends in the data. The enriched data can be stored in a database on fast secondary storage where it can be used repeatedly by many users to rapidly mine for phenomena of interest. Our research effort with SSM/I data shows that the approach gives anticipated results and has many potential applications in the mining of transient and long-term phenomena.

Introduction

Large tertiary-storage-based scientific data archives, such as NASA's Earth Observing System Data and Information System (EOSDIS), can potentially support many different data mining objectives. Our mining research has focused on developing the ADaM (Algorithm Development and Mining) data mining system (Hinke *et al.* 97) to mine Earth science data for phenomena relevant to the Earth science community. While not yet released as a production system, we have used ADaM to mine for various phenomena, ranging from very short duration lightning to longer duration mesoscale convective systems (large storms). ADaM has been used in our university laboratory and at NASA's Global Hydrology and Climate Center.

An impediment to the widespread application of data mining in scientific archives is the long duration of the data transfer from tertiary storage to memory (where it can be mined) and contention for tertiary-storage access (since the primary mission of these archives is filling data orders). To reduce this bottleneck, this research has focused on techniques for performing an initial mining of the archive for enriched data, whose volume is significantly smaller than the original data. This enriched data, however, preserves sufficient properties of the original data that it can be mined for specific phenomena of interest.

Enriched data could form a permanent, albeit significantly smaller, archive that could be stored in a database on secondary storage, rather than tertiary storage. Many users could quickly mine this enriched data for their particular phenomena of interest. The challenge, therefore, is to achieve substantial data reduction while ensuring that enriched data is general enough to support a wide range of phenomena mining objectives.

The approach described is analogous to the approach currently used on the world-wide-web to search for sites of interest. The web search engines have databases that hold information produced by web-robots crawling over various Internet web sites and capturing keywords of interest. A user performing a search using a search engine is accessing only the database created by web-robots. The approach described in this paper accomplishes a similar task for scientific data mining applications.

Creation of Enriched Data

The SSM/I (Special Sensor Microwave/Imager) data is used to support the research effort mining for enriched data. The SSM/I data, which is of great interest to Earth Scientists, is captured by sensors on near-polar-orbiting Defense Meteorological Satellites. These satellites are sun synchronous and pass over most parts of the Earth each day. This data has been used by scientists for detecting rain, mesoscale convective systems (severe storms), cloud liquid water, water vapor and ocean wind speed and for many other applications. (Devlin 95, Remote Sensing Systems 97) We have previously mined SSM/I data (with our ADaM data mining system) using various target-dependent mining algorithms for locating phenomena such as rain and mesoscale convective systems. (Hinke *et al.* 97)

Phenomena that represent short lived transients and long term trends are of interest to many users of the data. As a

This research was supported by a NASA Headquarters Grant NAGW .

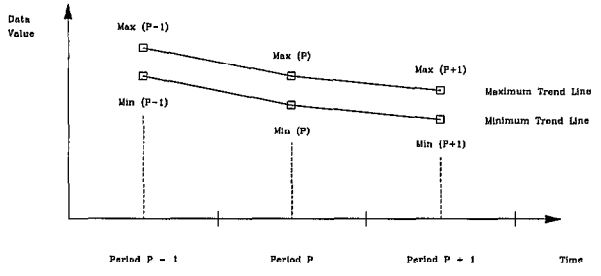
result the enriched data must include data that represents both types of phenomena. The next two sections describe the approach for extracting the desired enriched data, and the organization and storage of enriched data, respectively.

Approach

We assume that the surface of the Earth has been overlaid with a grid consisting of a number of equal degree bins, each representing a small area of the Earth's surface. Within remotely sensed data, a transient causes data to deviate significantly from the expected normal values. Two different methods were used to compute normal values for each bin. The first method approximated the normal value of a bin for a given day as the average of all data values recorded within the bin during a period of N days centered about the day of interest. The standard deviation of data values is used to determine values that deviate significantly from the normal.

In the second method, for each bin B, two trend lines are computed. For this purpose each year is partitioned into several periods of equal duration. For each bin the maximum and minimum values of data are found for each day of the period. The median of all maximum values of a period is considered as the normal maximum value for the day that marks the middle of the period. The line joining the normal maximum values computed for period P and period P+1 is taken as the normal trend of maximum value for the period spanned by the line. This is illustrated in Figure 2.1. The trend line of minimum value is determined for each bin in a similar fashion.

Figure 2.1 Trend Line Computation



Data values that are significantly above the maximum trend line or below the minimum trend represent transient events and become part of the enriched data. When tested through simulation the second method yielded greater data compression for the same level of mining accuracy than the first method. This is mainly due to the fact that the mean is more sensitive to the presence of transient events than the median. Therefore, the first method was discarded in favor of the second method.

Some sensors produce significantly different values over different types of terrain even when similar atmospheric conditions exist. For example, the SSM/I sensor response over water differs from its response over land under identical conditions. We used a relatively small bin size of one-half degree by one-half degree in order reduce the effects of terrain-type discontinuities. If a bin is large the possibility of including different types of terrain increases. This in turn may yield data values within the same bin that differ significantly from one another. Using trend lines for maximum as well as minimum values further reduces the effects of non-homogeneous bins.

Enriched Data

The enriched data obtained by phenomena independent mining consists of a transient component and a trend component. The transient component consists of all data values that are above the maximum trend line or below the minimum trend line by more than a specified threshold. The mining algorithms developed to operate on the original data can be used to mine the transient component. A measure of loss of accuracy due to data enrichment can be obtained by comparing mining results obtained from the original data to those obtained from transient component. As will be shown in the next section, for SSM/I data a 98% data compression is achieved, while losing only about 8% of data that represented mesoscale convective systems. As can be expected, as the compression rate is decreased, the number of lost data points also decreased. A further reduction in data may be achieved by using clustering and vector quantization techniques on enriched data.

A trend component is also saved for each bin for each period. For each period, the maximum values associated with various bins are stored in an array called MAX. Similarly, MIN holds minimum values associated with all bins for that period. The trend lines run from the middle of one period to the middle of the next and previous periods. Each value in the MIN or MAX array serves as the end point for two trend lines associated with consecutive periods. The trend lines characterize normal values of data for various bins and periods in a compact form. One may consider trend lines as effective means of representing all the data that is not included in the enriched transient data. For SSM/I data, trend line storage required less than 0.5 percent of the storage needed for raw data.

Target-Dependent Mining

Target-dependent mining using enriched data is illustrated with the help of the following two examples. The first example demonstrates mining of the transient component and the second example illustrates mining of the trend component of the enriched data.

Mesoscale Convective Systems

Consider the problem of mining for mesoscale convective systems(MCS) in SSM/I data. The algorithm used for mining for MCSs (Devlin 95) computes a polarization corrected temperature from two of the SSM/I channel values for each point vector. It then searches for a contiguous region of temperatures such that at least one

temperature value is below 225 degrees Kelvin and the remaining temperature values are below 250 degrees Kelvin. The contiguous region containing these values must represent an area of at least 2000 square kilometers.

Both the original and enriched data are mined for MCSs. The two results are used to assess the accuracy of phenomena-independent-mining. The experimental results provide an indication of how well the new approach preserves data needed for mining the MCS phenomena and the level of compression that can be accomplished.

Figure 3.1 Compression vs. Data Loss

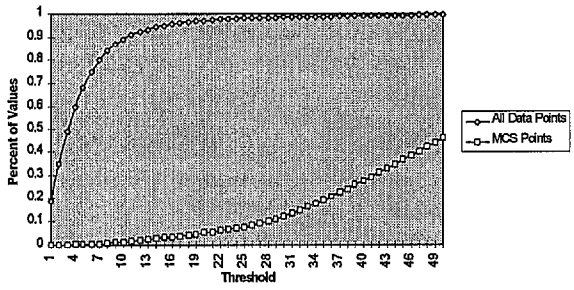


Figure 3.1 relates percent data compression to percent loss of MCS data. The horizontal coordinate represents the deviation threshold from normal values as defined by trend lines. The curve labeled "All Data Points" is a plot of the number of original data values in percent as a function of deviation threshold. For a given deviation threshold δ, it specifies the percentage of values that lie in those bins in which no value deviates from the trend lines by more than δ. This curve provides a quick estimate of the minimum compression that can be achieved using any given threshold. The actual compression will be slightly greater than our estimate since only the actual vectors that deviate from normal will be selected instead of all vectors in the bin. The curve labeled "MCS Points" is a plot of the number of MCS data points in percent, as a function of deviation threshold.

By using a deviation threshold of 15 degrees, a compression ratio of at least 95% is achieved. Next, vectors of interest were selected from the original data which were either at least 15 degrees warmer than the maximum trend line or 15 degrees cooler than the minimum trend line. This resulted in the selection of 2.12% of the original data vectors, yielding an actual compression of 97.88%. Both the original data set and the enriched subset were then used to mine for MCSs, and the results were compared. It was found that 92.25% of the MCS values present in the original data were still identifiable in the enriched data. This means that a very large percentage of the values composing these transient phenomena are present in a very small percentage of the raw data. The algorithm drops some MCS values when size requirement was not satisfied due to one or two missing MCS values. As a result of this, the percent of missing MCS values appears larger than the actual number. We are looking at ways to address this problem, by perhaps, including values that are adjacent to deviation values in the enriched data. This will decrease the compression but may improve results. Finally, it should be noted that the MCS data represents only 6.7% of the enriched data.

Trend Mining

This section presents examples of data mining using the trend component of the enriched data. As previously described, phenomena-independent mining produces two trend lines for each bin and for each time. Let max(B,P) be the maximum trend line of bin B for the time period P and min(B,P) be the minimum trend line. Each of these trend lines are defined by the pair (I,S), where I[B,P] represents the intercept and S[B,P] represents the slope of the trend line for bin B and period P.

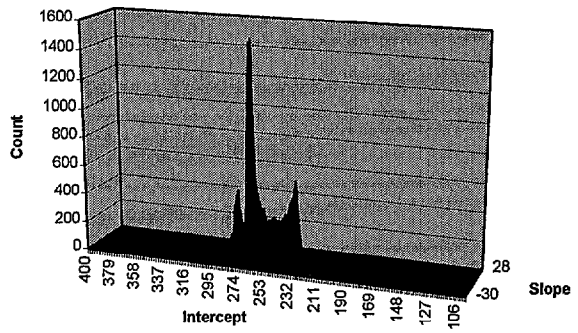
Consider a 2-D space in which the vertical coordinate represents I[B,P] and the horizontal coordinate represents S[B,P]. This space, referred to as the slope-intercept space, is also used in Hough transforms for image processing (Gonzalez & Woods 1992). It is obvious that each trend line maps to a single point in the slope-intercept space, and that all identical trend lines map to the same point. Let (Imin, Imax) be the range of intercept and (Smin, Smax) be the range of slope. For digital implementation, the slope-intercept space defined by these ranges needs to be quantized. The temperature values in SSM/I data that were used in our work lie between 100 and 400 degrees. Also, the temperature does not change by more than 30 degrees during any single time period. Selection of one-degree resolution for the intercept (temperature) and one degree for slope (change in temperature) resulted in a 300X60 slope-intercept array which is denoted by A.

Assume that the value of each cell in the slope-intercept array is initialized to zero. The line's slope and intercept determine the cell to which a trend line maps. Each time a line falls into a cell, the value of the cell is incremented. After mapping all trend lines, if A(S, I) = N then it can be concluded that N bins are characterized by trend lines with slope = S and intercept = I. The sum of all elements of A is equal to the total number of bins. Note that one slope-intercept array is needed for each time period. Although the actual location of the bins that contributed to the count can not be determined from this array, they can be easily retrieved from the target-independent trend data.

The slope-intercept array provides useful global information for the time period that it represents. The array A can also be used as an image for visual presentation of the general temperature trend for the time period for which values are included. An example of such an image is shown in Figure 3.2. The high peak represents a dominant mode for this data. Using a cutoff count of 600, the target-independent trend results can be rapidly searched to find the bins to which this peak corresponds. The results of this

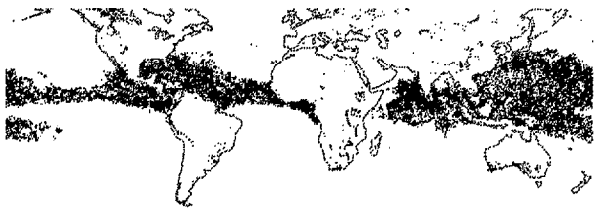
search are displayed on a map image shown in Figure 3.3.

Figure 3.2 Slope Intercept Array



Another mining application for the slope-intercept array is finding the number of bins that experienced more than 20 degrees gain in temperature. Also, the presence or absence of bins with a specified trend in a given period of time can be determined with negligible computation.

Figure 3.3 Dominant Mode from Slope Intercept



There are also other possible ways of mining the trend lines. The trend lines can be plotted over extended periods of time for a given region to study how the climate has changed. Trend lines from several regions can be compared to find areas with similar climate, or to find boundaries between climate zones. Finally, the trend line endpoints for adjacent months can be stored as images and visualized in several ways. For instance, a single set of endpoints forms an image that indicates normal values for a given period of time. A sequence of such images can be used to animate the climate changes with season. These images can be displayed either with the actual values or as difference images where the trend between frames (which is the slope of the trend line) is displayed.

Related Work

Researchers, such as (Berndt & Clifford 96) mine trend data for patterns of interest, such as technical analysis patterns that apply to the stock market. This type of mining could be applied to finding patterns of interest in long term trends constructed by piecing together the bin-specific trends described in this research. Our concept of deviation is related to measures of interestingness (Silberschatz & Tuzhilin, 95). A group from UCLA and JPL have developed a system called Oasis for performing data mining of geophysical data. (Mesrobian, et. al. 1996). This work does not address the issues raised in this paper.

Conclusions

In this paper we have presented a mining approach called target independent mining that is capable of accomplishing significant data compression while retaining almost all information needed for the mining of transient phenomena and long term trends in large scientific data. Experiments using SSM/I data have shown that almost 98 percent compression in data is achieved by losing less than 8 percent of useful transient data for the MCS phenomena. The availability of enriched data on secondary storage or in a database which allows many users with different mining objectives to mine rapidly is a significant step in the right direction as compared to mining data stored on tertiary storage. While the reported work is still in the research stage, we are looking for opportunities to field a system using the approaches presented in this paper.

References

Berndt, D.J and Clifford, J. Finding Patterns in Time Series: A Dynamic Programming Approach. *Advances in Knowledge Discovery and Data Mining*, eds. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy. AAAI Press/ The MIT Press, 1996.

Devlin, K.I. Application of the 85 Ghz Ice Scattering Signature to a Global Study of Mesoscale Convective Systems. Master's thesis, Meteorology, Texas A&M University, August 1995.

Gonzalez, R.C. and Woods, R.E. *Digital Image Processing*. Reading, Mass.; Addison-Wesley.

Hinke, T.H.; Rushing, J.; Kansal, S.; Graves, S.; Ranganath, H.; and Criswell, E. Eureka Phenomena Discovery and Phenomena Mining System. *Proceedings: 13th International Conference on Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography and Hydrology*, February 1997.

Mesrobian, E.; Muntz, R.; Shek, E.; Nittel, S.; La Rouche, M., Kriguer, M.; Mechoso, C.;, Farrara, J; Stolorz, P.; and Nakamura, H. Mining Geophysical Data for Knowledge. *IEEE Expert*, October 1996.

Remote Sensing Systems, Data Products, URL http://www.ssmi.com.

Silberschatz, A. and Tuzhilin, A. On Subjective Measures of Interestingness in Knowledge Discovery. *Proceedings of the First International Conference on Knowledge Discovery & Data Mining*, August 1995.