Refining query previews techniques for data with multivalued attributes: The case of NASA EOSDIS

Catherine Plaisant, Maya Venkatraman,
Kawin Ngamkajornwiwat

Human-Computer Interaction Laboratory
University of Maryland
Institute for Advanced Computer Studies
College Park, MD 20742
plaisant@cs.umd.edu
http://www.cs.umd.edu/hcil/eosdis

Randy Barth, Bob Harberts, Wenlan Feng

Raytheon Corporation
4500 Forbes Blvd.
Lanham, MD – 20706

Robert.Harberts@gsfc.nasa.gov

Abstract

Query Previews allow users to rapidly gain an understanding of the content and scope of a digital data collection. These previews present overviews of abstracted metadata enabling users to rapidly and dynamically avoid undesired data. In this paper we present our recent work on developing query previews for a variety of NASA EOSDIS situations. We focus on approaches that successfully address the challenge of multi-valued attribute data. Memory requirements and processing time associated with running these new solutions remain independent of the number of records in the dataset. We describe two techniques and their respective prototypes used to preview NASA earth science data.

1. Introduction

Soon, large numbers of users from varying background will visit the NASA EOSDIS information systems to locate earth science data of interest. Such information systems must accommodate different levels of experience with the content area, with the information system itself, and with information seeking approaches in general [Marchionini, 95]. The traditional approach to querying is to use a form fill-in interface, but such an approach leads to user frustration when the query returns either zero hits or a very large number of hits. Often, users cannot estimate the total number of hits their query actually returns because the system truncates to the first 25 - 50 hits. It is difficult to estimate how much data is available on a given topic and how to increase or reduce result set sizes. b

Our approach to overcome these challenges involves the presentation of overviews and previews of abstracted metadata that allow users to perform rapid and dynamic elimination of undesired data. The reduced volume of the abstracted metadata allows queries to be previewed and refined locally by the user before they are submitted over the network.

We developed the concept of query previews [Doan, 96; Plaisant et al, 98] that allows users to:

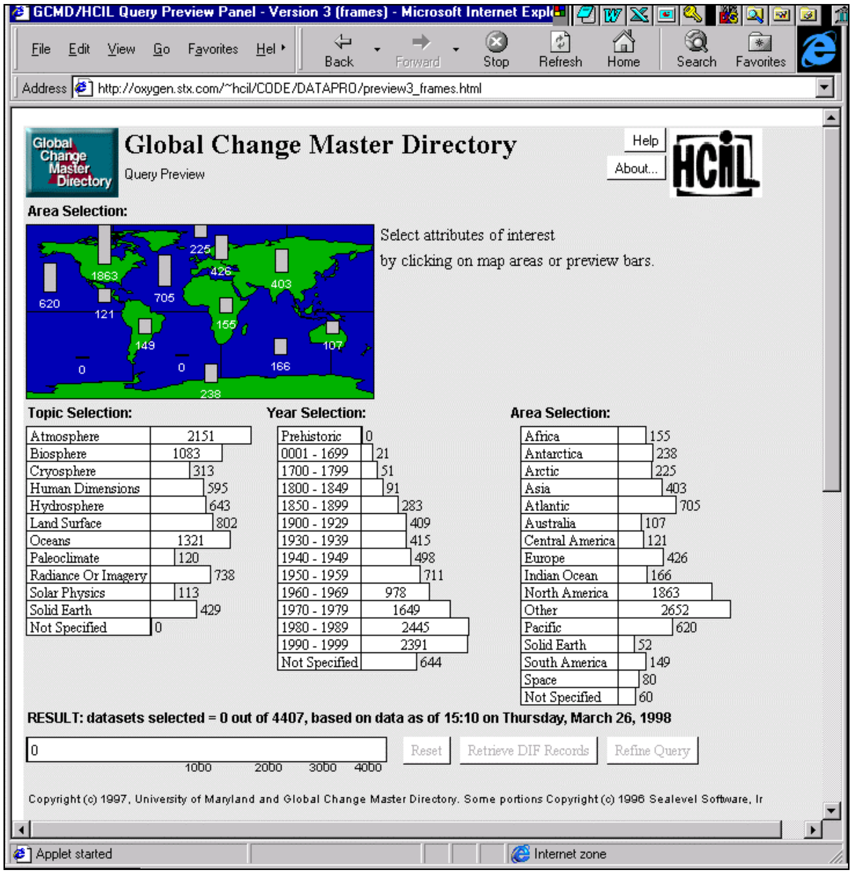
- rapidly gain an understanding of the content and scope of a digital library,
- dynamically specify their initial query,
- review the effect of their query on the expected total number of hits, and
- adjust the query accordingly in a dynamic query style of interaction [Shneiderman, 94]

2. Current EOSDIS use of query preview

After our early prototypes [Doan, 96], we worked in close collaboration with the Global Change Master Directory (GCMD) team to apply our techniques to an operational system [Greene, 98]. The GCMD serves as catalog or directory for all available earth science data. The GCMD is a sort of "Yellow Pages" for Earth Science Data. Currently there are around 5000 datasets available in the directory.

Our GCMD query preview interface (Figure 1a) features a single-screen overview of all the datasets in GCMD, and allows users to rapidly narrow down the number of datasets according to their interest. Three high-level attributes (location, years of coverage, and general topic area) are used to define the users envelope of interest. These three attributes were chosen by the GCMD staff as the most salient ones for a majority of users. The large set of attribute values for each of the three attributes was aggregated into approximately ten high-level attribute designations. The granularity of the attributes was deliberately kept rather crude (or large), in order to be able to represent the data in a single overview screen.

HCIL Technical Report No. 98-17 (April 1999); to appear in Proceedings of ADL'99, IEEE Computer Press
http://www.cs.umd.edu/hcil



Global Change Master Directory

Query Preview

Area Selection:

Select attributes of interest

by clicking on map areas or preview bars.

Topic Selection:	
Atmosphere	2151
Biosphere	1083
Cryosphere	313
Human Dimensions	595
Hydrosphere	643
Land Surface	802
Oceans	1321
Paleoclimate	120
Radiance Or Imagery	738
Solar Physics	113
Solid Earth	429
Not Specified	0

Year Selection:	
Prehistoric	0
0001 - 1699	21
1700 - 1799	51
1800 - 1849	91
1850 - 1899	283
1900 - 1929	409
1930 - 1939	415
1940 - 1949	498
1950 - 1959	711
1960 - 1969	978
1970 - 1979	1649
1980 - 1989	2445
1990 - 1999	2391
Not Specified	644

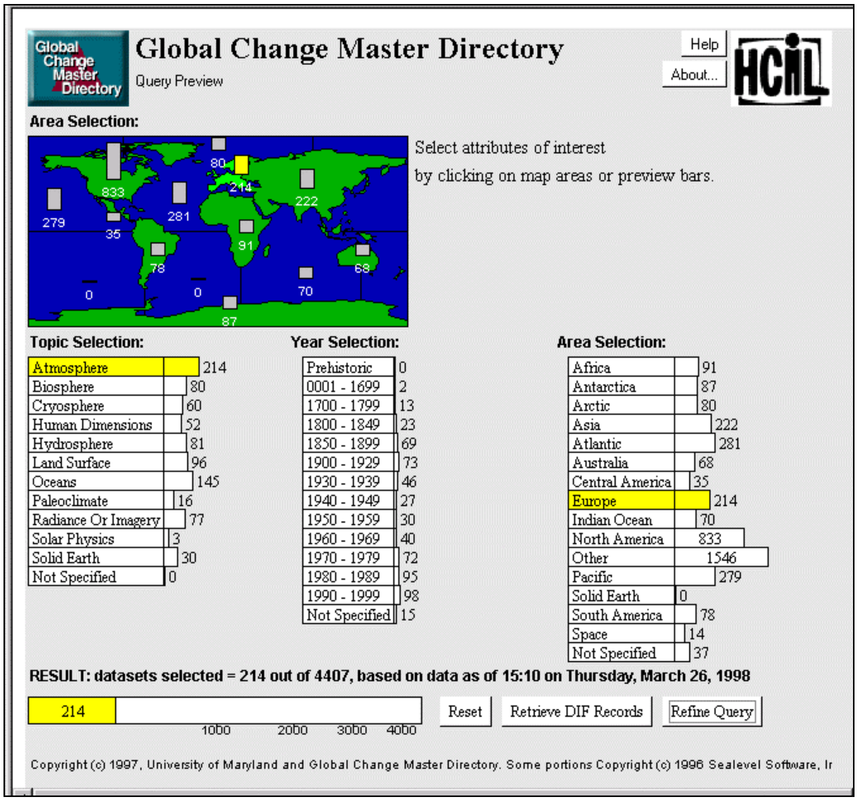
Area Selection:	
Africa	155
Antarctica	238
Arctic	225
Asia	403
Atlantic	705
Australia	107
Central America	121
Europe	426
Indian Ocean	166
North America	1863
Other	2652
Pacific	620
Solid Earth	52
South America	149
Space	80
Not Specified	60

RESULT: datasets selected = 0 out of 4407, based on data as of 15:10 on Thursday, March 26, 1998

0

Copyright (c) 1997, University of Maryland and Global Change Master Directory. Some portions Copyright (c) 1996 Sealevel Software, Ir

Figure 1a. The interface for the NASA Global Change Master Directory provides a single screen overview of the digital library.



Global Change Master Directory

Query Preview

Area Selection:

Select attributes of interest

by clicking on map areas or preview bars.

Topic Selection:	
Atmosphere	214
Biosphere	80
Cryosphere	60
Human Dimensions	52
Hydrosphere	81
Land Surface	96
Oceans	145
Paleoclimate	16
Radiance Or Imagery	77
Solar Physics	3
Solid Earth	30
Not Specified	0

Year Selection:	
Prehistoric	0
0001 - 1699	2
1700 - 1799	13
1800 - 1849	23
1850 - 1899	69
1900 - 1929	73
1930 - 1939	46
1940 - 1949	27
1950 - 1959	30
1960 - 1969	40
1970 - 1979	72
1980 - 1989	95
1990 - 1999	98
Not Specified	15

Area Selection:	
Africa	91
Antarctica	87
Arctic	80
Asia	222
Atlantic	281
Australia	68
Central America	35
Europe	214
Indian Ocean	70
North America	833
Other	1546
Pacific	279
Solid Earth	0
South America	78
Space	14
Not Specified	37

RESULT: datasets selected = 214 out of 4407, based on data as of 15:10 on Thursday, March 26, 1998

214

Copyright (c) 1997, University of Maryland and Global Change Master Directory. Some portions Copyright (c) 1996 Sealevel Software, Ir

Figure 1b. When users select attribute values (here Atmosphere and Europe), all counts are updated using pre-computed preview tables. This gives users an idea of the size and distribution of the result set before the query is submitted, and reduces the zero-hit or mega-hit query problem. Following the principles of dynamic queries, the preview bars are updated immediately (in less than 100msec.). The result bar at the bottom shows the total number of selected data sets.

Our overall finding was that the query preview interface concept is a feasible solution in the EOSDIS directory environment. Consensus was reached rapidly on attributes and values selection. Performance was acceptable and continued to improve as speed issues were solved. Our experience also confirmed the importance of metadata accuracy and completeness. The query preview interface makes visible any problems or holes in the metadata that are unnoticeable with classic form fill-in interfaces (see for example the large number of datasets with unspecified year coverage). This could be seen as a problem for the interface but we think that in the long term it will have a beneficial effect on the quality of the metadata, as data providers will be compelled to produce more accurate and complete metadata.

This interface is included within the operational GCMD where it is offered as an alternative experimental service (http://gcmd.nasa.gov).

3. The challenge of data with multi-valued attributes

But the work with GCMD also revealed some problems with our early assumptions about the NASA data. Our original solution consisted of using as a preview table a simple N dimensional array for an N attribute preview table (e.g. for GCMD a small 12x14x16 table of counts). This technique works very well for data having a single value for each attribute (e.g. the case where each record only has one topic, and covers only one time period, and one area of the globe). However, EOSDIS datasets have multi-valued attributes that require the duplication of records in the N dimensional array preview tables (e.g. when datasets have multiple topics, or cover multiple areas). We had assumed that a small duplication factor would remain unnoticed by users but the real data revealed a 700% duplication factor that was very noticeable as the total number given in the preview was very wrong.

The obvious way to handle this problem would be to count the datasets in multiple cells of the table but keep track of all possible intersections. This would lead to preview tables and computation time growing exponentially with the number of attributes and attributes values, which in turn leads to high processing time that defeat the purpose of query previews.

This unexpected problem needed a rapid solution for GCMD, and led to an hybrid solution that keeps the list of dataset IDs with the dataset counts, so that all duplicates can be removed to calculate the total number of dataset accurately. This hybrid technique works well for the number of datasets kept at GCMD (i.e. in the thousands). It results in a long but still reasonable loading time (10-20 sec. for the applet and preview table) and good response time (less than a second) once the applet is loaded and users start making selections.

In summary:

- When the data is single-valued (i.e. the case of most relational database) the interface shown in Figure 1 is simple to implement however large the number of records is.
- When the data has multivalued attributes, the hybrid GCMD technique is acceptable for numbers of records up to 5,000-10,000 records.
- For a large number of records, alternative techniques have to be designed. In the case of EOSDIS such alternate techniques are needed to search within single dataset or within combined datasets that could include hundreds of thousands of records (called granules in the case of EOSDIS).

The following section discusses the solutions we developed to respond to this challenge.

4. Many practical solutions

At this stage of the research we do not know of any general solution to the problem that is applicable to datasets of any size. The size of the preview table and the computation time grows exponentially with the number of attributes and attribute values. Nevertheless, we know that there are many practical solutions:

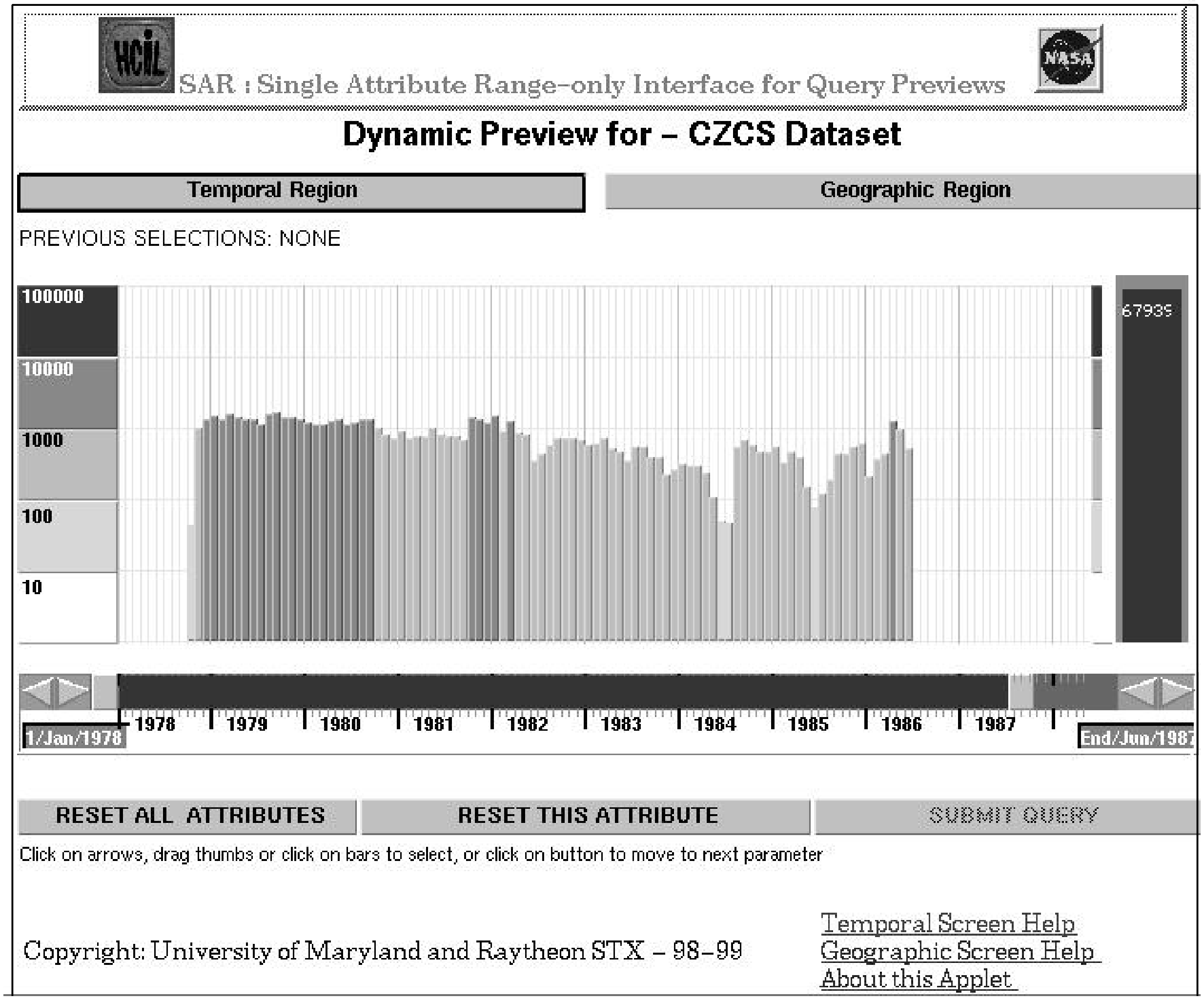
- For small number of records (like GCMD) a hybrid solution listing individual record IDs in the preview table works well.
- In all cases, we know that reducing the number of attributes, or reducing the query attributes values, or limiting the number of user selections per attribute dramatically reduces the size of the query preview table and can lead to acceptable size tables. The system architect will have to judge if the resolution or restrictions lead to still useful interface.
- For data having attributes with range values (e.g. temporal coverage is a range of dates) that lead to range queries, we can use specific algorithms that results in query preview tables whose size is independent of the number of the records.
- Finally it is possible to reduce the preview to a binary preview in which the users is informed of the existence of data but not of the volume of data available. In this case duplication does not matter.

We investigated the last 2 solutions (range queries and binary previews) and now have working prototypes of those techniques using EOSDIS datasets.

4.1 Approach 1 - Series of single-attribute, range queries

Range queries on range variables with multi-valued attributes are frequent. This is the case for many EOSDIS queries (e.g. temporal coverage, many geographical coverages, see Figure 2a&b) For those range queries we can use a special algorithm that was devised by our colleagues Richard Beigel and Egemen Tanin at HCIL to makes the preview tables independent of the number of records using Euler's formula [Beigel & Tanin, 1998].

Using Euler's formula it is possible to make query previews work without transferring actual data over the network (i.e. no need to transmit the record IDs). Euler's formula computes the number of datasets that actually fulfill a query by using a few simple arrays of data. For example, in the case of one-dimensional data (e.g. temporal data) two arrays are used. One array specifies the counts of datasets for each cell and the second array specifies the number of datasets that "cross-over" each horizontal wall of the cells. For two-dimensional data (e.g. geographic data) two more arrays are used. The next array specifies counts for the number of datasets that crossover the vertical walls and the final array



SAR : Single Attribute Range-only Interface for Query Previews

Dynamic Preview for – CZCS Dataset

PREVIOUS SELECTIONS: NONE

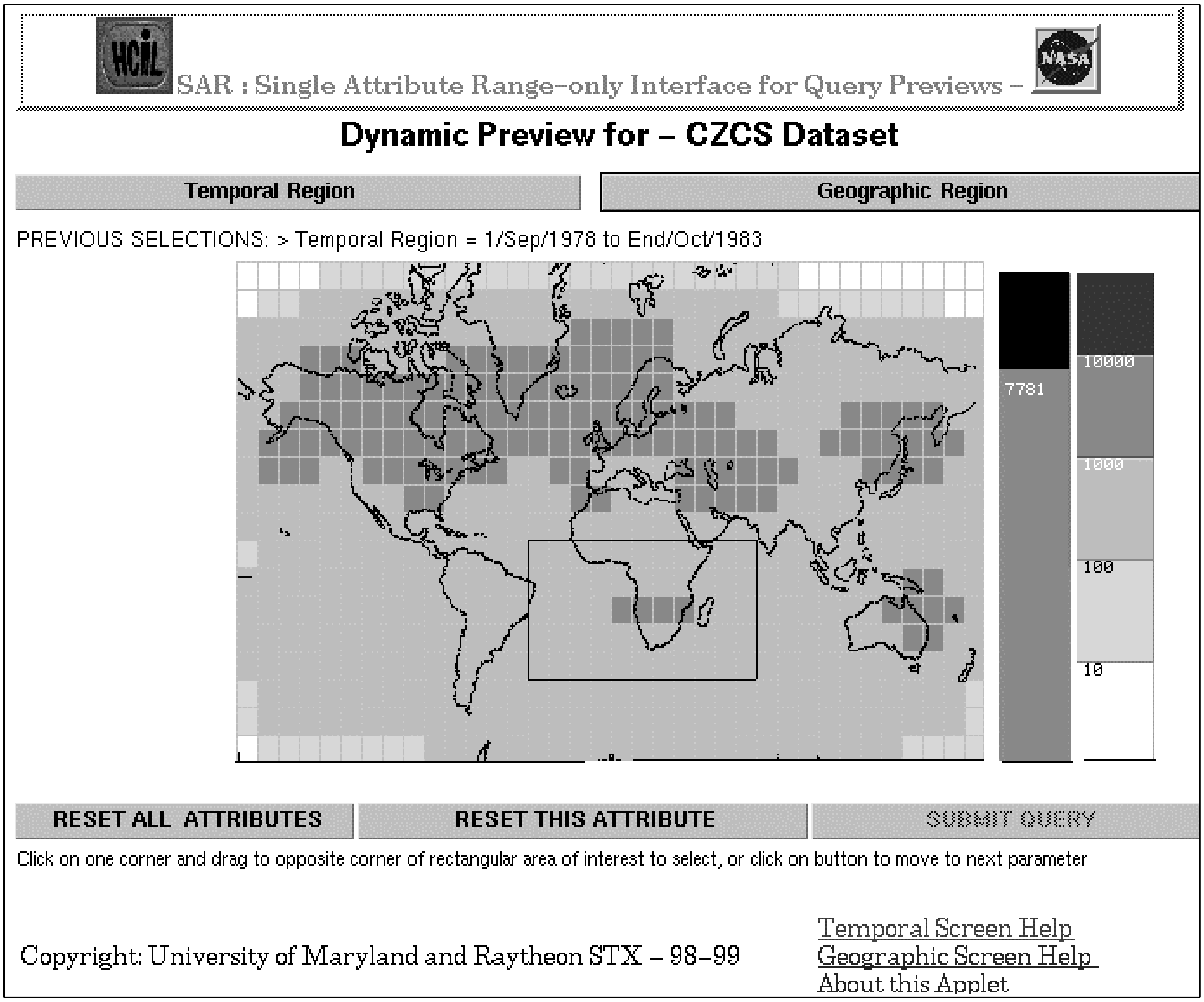
Click on arrows, drag thumbs or click on bars to select, or click on button to move to next parameter

Copyright: University of Maryland and Raytheon STX – 98-99

Figure 2a: This alternative approach presents a series of single attribute, range only, query previews. The first attribute shown here is Time. Each month has a bar showing the total number of records for that month. The scale is logarithmic so the taller bars correspond to much larger number of records. Color-coding reinforces the scale. We can see that data is available from Jan 1978 to Jun 1987. Around May 1984 the instrument must have had problems and therefore the number of granules is much smaller than usual. Using the sliders, users can select a time range and see the total number of granule selected on the large bar on the right. This approach scales up because the preview table size is independent of the number of records. This approach is also able to handle multivalued attribute data

specifies the counts for the number of datasets that cross over each corner or vertex of the cells.

To find the number of records that fall within a query space, simple additions and subtraction operations on arrays are the only computations required. The arrays are either pre-computed or calculated on the fly. The initial arrays for each attribute (when there are no selections yet) can be pre-calculated and stored. However, when the user makes a selection and moves to the next attribute, the relevant arrays have to be recalculated based on the previous query. For instance, if the temporal query was from May 1978 to May 1980, and the geographic attribute is selected, the four geographic arrays will have to be calculated using all the data that was collected between May 1978 and May 1980. This process is run on the server. We are investigating several alternatives. The first solution is to pass the query values to a program that searches a database of pre-bucketed data (each record has already been assigned to one or more bucket/cell of the query space), and pre-computes the arrays every time a query is made. This technique is simple but the delay will be proportional to the number of records. Another more general alternative is to implement a three-dimensional



SAR : Single Attribute Range–only Interface for Query Previews –

Dynamic Preview for – CZCS Dataset

PREVIOUS SELECTIONS: > Temporal Region = 1/Sep/1978 to End/Oct/1983

Click on one corner and drag to opposite corner of rectangular area of interest to select, or click on button to move to next parameter

Copyright: University of Maryland and Raytheon STX – 98–99

Figure 2b: When users select a new attribute, (in this case - geographical coverage region), the time range selected in Figure 2a is sent to the server, which then generates a new preview table. Only color coding is used on the map to show the range of numbers of records available. When users select a range of latitudes and longitudes (a rectangular box), it will be passed to the server to generate the next preview or return the results.

matrix of the counts (cell counts, overlap counts and vertex counts), so that the only on-the-fly processing is to extract a cube subset of this three-dimensional array. This technique is independent of the number of records and was found to be quite fast when saving the datacube in a database.

In the case of the example given in Figure 2a&b, the table containing the datacube has 1348 records for the Coastal Zone Color Scanner (CZCS) database with ten degree geographic resolution and 2600 records for the CZCS database with five degrees geographic resolution. It takes about 5 - 10 seconds to load the applet and the initial preview table. When users change attribute, the table is recomputed on the server and reloaded in about 3-5 seconds.

When comparing this technique with the hybrid technique used for GCMD, readers should remember that the Single Attribute - Range (SAR) method is completely independent of the number of records, therefore scalable to any data collection. In addition, with the SAR interface we were able to achieve a much finer granularity for the attribute value selection (e.g. small 5-degree grid cells instead of whole continents for the map). On the other hand, only one attribute is being previewed at a time - instead of three for the GCMD example, and only range queries are possible.

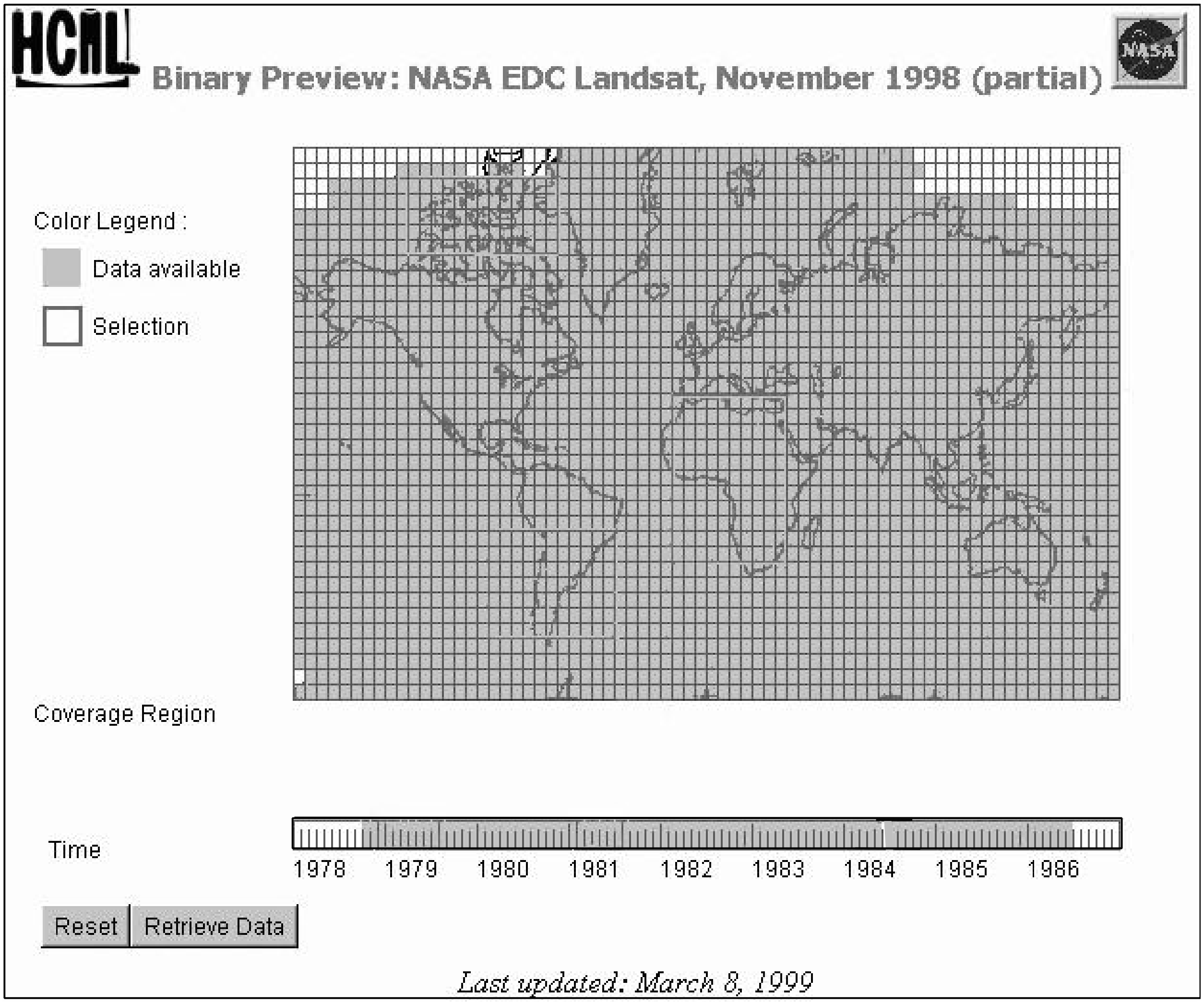
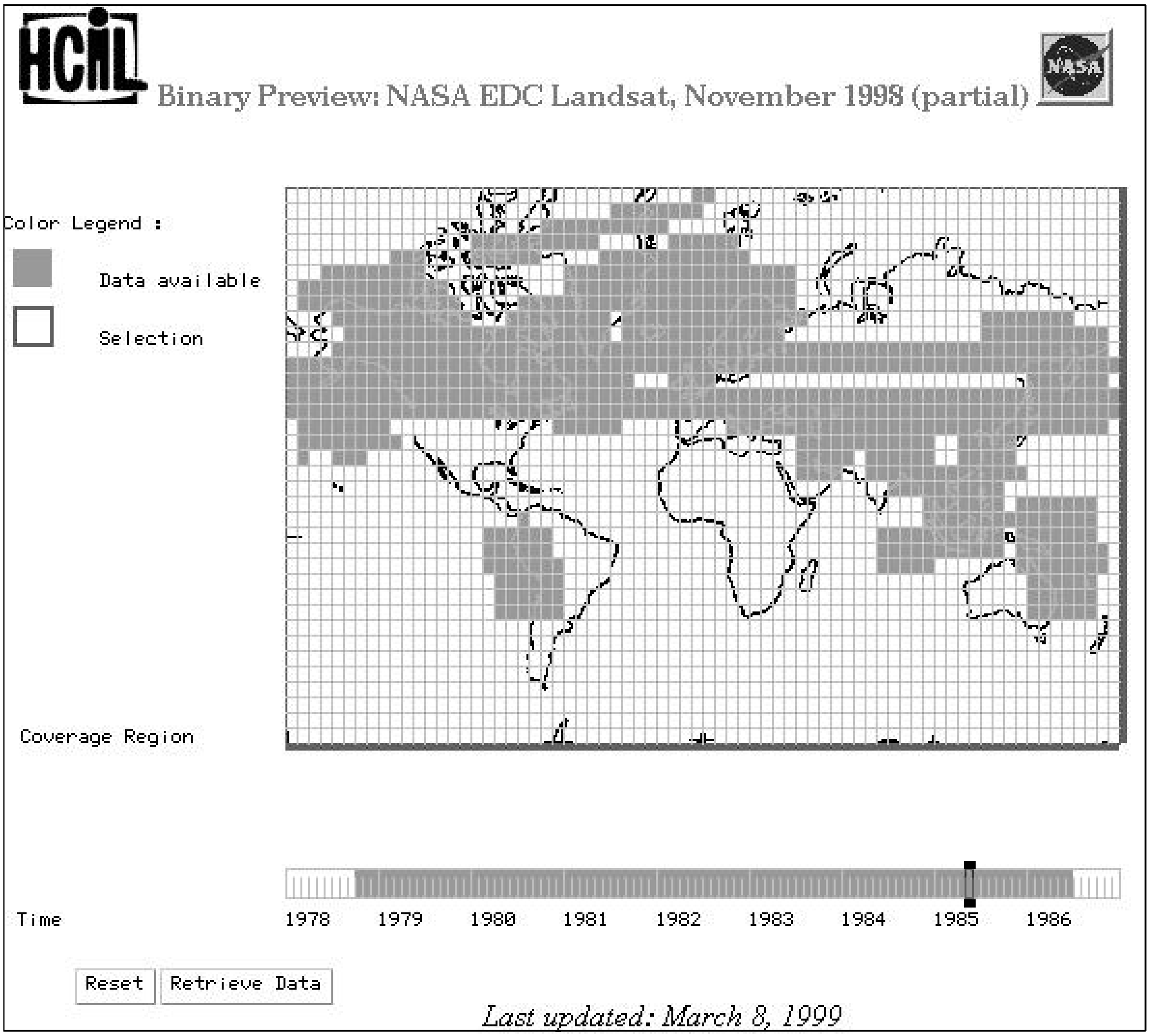


Figure 3a: At first the preview shows an overview of the dataset (here about 60,000 granules from the Coastal Zone Coastal Scanner dataset from the NASA Goddard data center). It shows that almost the whole globe was covered from about mid 1978 to mid 1986 but no data was collected after mid 1986.



Figures 3b: As users select one month, the binary previews on the map display the data available for those months, revealing that only a subset of the globe was actually observed during that month. Users can select more that one-month at a time. Similarly users can select an area(s) on the map and see during what year and months the data was actually collected for that area.

4.2 Approach 2: Binary previews

Binary previews used a different approach. Previews do not provide counts but merely indicate the presence or absence of data. Series of small binary masks can be compressed and transferred quickly to the applet. These masks can be combined (OR-ed together) when users make selections. Figure 3a shows that the records cover most of the world, and several years, but as users select attribute values (e.g., a month in Figure 3b) they can see that the data is actually pretty sparse since each month has many coverage holes. Similarly users could select an area of the world and see what is the time coverage for that area. Selections could be ranges as well as sets of disjoined grid cells. Because the masks are small, several attributes can be presented at once, providing a very interactive exploration of the data space. Unlike the first approach, there is no need for recalculation at the server. In summary, binary previews present a much simpler approach to query previews, at the cost of not providing counts. Zero-hit queries can still be reduced, but not the mega-hit queries. Binary previews are particularly appropriate when the data is sparse or has major "holes". This is the case for many EOSDIS datasets where instruments are not always active, and the data is not always captured and stored in a reliable fashion.

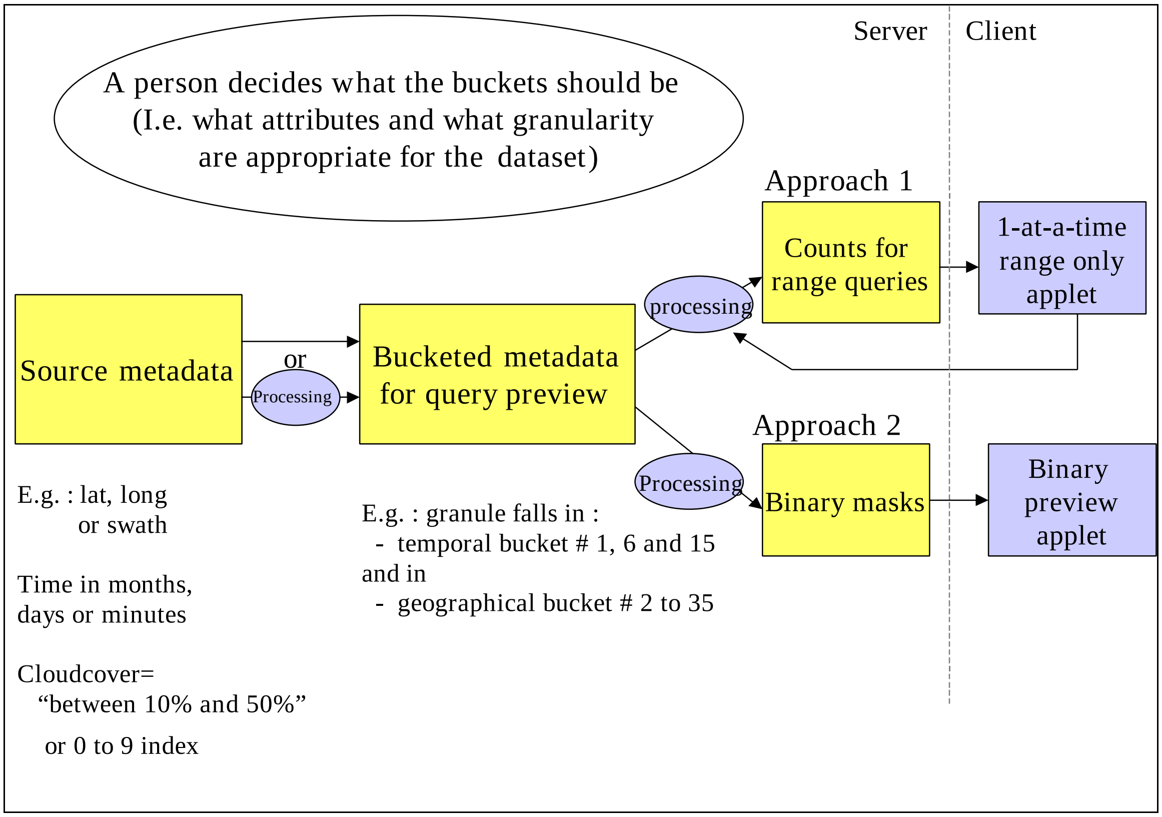

Server
Client
A person decides what the buckets should be (I.e. what attributes and what granularity are appropriate for the dataset)
Approach 1
Counts for range queries
1-at-a-time range only applet
processing
Source metadata
or
Processing
Bucketed metadata for query preview
Approach 2
Processing
Binary masks
Binary preview applet
E.g. : lat, long or swath
Time in months, days or minutes
Cloudcover= "between 10% and 50%"
or 0 to 9 index
E.g. : granule falls in :
- temporal bucket # 1, 6 and 15
and in
- geographical bucket # 2 to 35

Figure 4: Schematic diagram of the meta-data pipeline.

5. Metadata pineline

Figure 4 illustrates the process used to generate data that is visualized by the applets. Attributes for the previews and their granularity (bucket sizes) are determined by designers (after taking into account user requirements), in the initial phase of design. The source meta-data is then processed to generate the bucketed metadata, using these attributes and granularity values as specifications for the output. This intermediate data is used to generate the masks for the binary interfaces and the data cube for the SAR interface. Both the applets communicate with the server through a customized communications package. The binary applet communicates with the server only once when it retrieves the set of masks. This retrieval occurs when the applet is loaded. The server returns preview arrays that are entirely composed of zeros and ones for each of the display elements or attributes on the binary screen. The SAR interface sends a query to the server every time the user moves from one attribute screen to the other, or when the user resets all the previous selection. The server returns two arrays for a temporal query and four arrays for a spatial query. The server is multi-threaded and can handle multiple users at a time.

6. Related work

An early proposal for volume previews in a database search is described in [Heppe et al. 85]. The "Dining out in Carlton" example was provided to illustrate a search technique (for a specific restaurant) based on the volume preview of the number of the available restaurants. However, query previews were not exploited to support dynamic queries. Retrieval by reformulation is a method that supports incremental query formation by building on query results. Rabbit [Williams 84] and Helgon [Fisher et al. 89] are examples of retrieval systems based on the retrieval by reformulation paradigm. INQUERY uses a ranked output information retrieval system for a library catalog containing about 300,000 documents [Veerasami, 95]. Its interface supports a visualization scheme that illustrates how the query results are related to the query words. Tilebars, visualizes term distribution information in each document to supplement result lists in full text retrieval systems [Hearst, 95]. Finally datacubes [Rousopoulos, 97] seems a promising data structure that may impact the generation of query preview tables.

7. Future work

If enough quantitative guidelines can be devised we envision a tool that would automatically examine the data and guide designers toward an appropriate query preview interface. Such a tool would analyze the attributes of the data and suggest candidates for the preview. For example location should not be used for previewing if all records are global and cover the whole earth. On the other hand, an attribute such as cloud cover is good for previews since it has ordinal values and often has only a single value per record. Such a query preview advisor (or possibly later on “generator’) would provide useful assistance to data producers and data centers and facilitate the inclusion of query previews in a variety of applications.

8. Conclusions

This paper presents an update on our work developing query previews for a variety of EOSDIS situations. We focused here on approaches that successfully address the challenge of multi-valued attribute data while remaining independent of the number of records. We proposed two techniques and showed examples of their use with NASA data. We are still in the process of refining our software and evaluating its performance but the two techniques are promising. The binary previews may well be the easiest one to implement first in a large operational context such as EOSDIS because it does not require on the fly re-processing at the server and works quickly with a small applet. The other approach, single attributes - range only is more complex to implement, but provides the important counts that help contain the mega-hit problem. Both approaches scale up to collections of any size. More details on the comparisons between the two techniques have been presented in Table 1. Query previews have been very well received by users and we believe those new techniques will broaden their domain of application.

--- New proposed techniques ---		**--- Original techniques ---**	
Series of single attribute range queries	**Binary preview**	**Hybrid query preview** (GCMD prototype) List of IDs kept with the counts	**Initial query preview technique** (N dimensional array)
Allow multi-valued attributes			Single valued attributes only
Gives counts	No counts	Gives Counts	
Reduces zero-hit problem AND Reduces mega-hit problem	Reduces zero-hit problem	Reduces zero-hit problem AND Reduces mega-hit problem	Reduces zero-hit problem AND Reduces mega-hit problem
Dynamic on-the-fly processing required At server side	No reprocessing		
Preview table size independent of number of records		Preview table size a function of the number of records	Preview table size independent of number of records

Table 1: Informal comparison between the different options

Acknowledgements

We want to thank Egemen Tanin for his help with the range query algorithm and Ben Shneiderman for his guidance and reviews. This work is supported in part by NASA ESDIS, NAG52895.

References

Beigel, R., Tanin, E., The Geometry of Browsing, the 3rd *Latin American Symposium on Theoretical Informatics*, 1998. www.cs.umd.edu/hcil/eosdis

Heppe, D. L., Edmondson, W. H. and Spence, R., Helping both the novice and advanced user in menu-driven information retrieval systems, *Proc. of British HCI85 Conf.*, 1985., pages 92-101.

Doan, K., Plaisant, C. and Shneiderman, B., Query previews in networked information systems, *Proc. of the Forum on Advances in Digital Libraries*, IEEE Computer Society Press, 1996, pages 120-129.

Fischer and Nieper-Lemke, H., HELGON: Extending the retrieval by reformulation paradigm, *Proc. of ACM CHI'89 Conf.* , 1989, pages 357-362.

Greene, S., Tanin, E., Plaisant, C., Shneiderman, B., Olsen, L., Major, G., Johns, S., The End of Zero-Hit Queries: Query Previews for NASA's Global Change Master Directory, *University of Maryland Computer Science Technical Report*, CS-TR-3855 - www.cs.umd.edu/hcil (Dec. 1997).

Hearst, M., Tilebars: Visualization of term distribution information in full text information access, *Proc. of ACM CHI 95 Conf.* , Denver CO, 1995, pages 59-66.

Marchionini. G., *Information Seeking in Electronic Environments*. Cambridge University Press, UK, 1995.

Plaisant, C., Bruns, T., Doan, K. and Shneiderman, Interface and Data Architecture for Query Preview in Networked Information Systems, to appear in ACM TOIS as a Practice and Experience Paper.

Roussopolos, N., Kotidi, Y., Roussopolos, M., Cubetree: organization of and bulk incremental updates on data cube, *Proc. ACM SIGMOD 97.*

Shneiderman, B., Dynamic queries for visual information seeking, *IEEE Software 11*, 6, 1994, pages 70-77.

Veerasamy, A. and Navathe, S., Querying, navigating and visualizing a digital library catalog, *Proc. of the Second International Conf. on the Theory and Practice of Digital Libraries*, 1995 (URL: http://www.csdl.tamu.edu/DL95/)

Williams, M.D.,What makes RABBIT run?, *International Journal of Man-Machine Studies 21*, 1984, pages 333-335.