VOMegaPlot: efficient plotting of large VOTable datasets

Nilesh Urunkar[1], Ajit K. Kembhavi[2], Ameya Navelkar[1], Jagruti Pandya[1], Vivekananda Moosani[1], Prameela Nair[1] and Mohasin Shaikh[2]

[1]Persistent Systems Pvt. Ltd., Bhageerath 402, Senapati Bapat Road, Pune 411016, India
email: urunkar,amey,jagruti_pandya,vivekananda_moosani,prameela_nair, shaikh@persistent.co.in

[2]Inter-University Centre for Astronomy and Astrophysics (IUCAA),
Post Bag 4, Ganeshkhind, Pune University Campus, Pune 411 007, India
email: akk@iucaa.ernet.in

Most plotting tools tend to load all the data to be plotted into main memory and then use the in-memory data for the actual operations such as plotting. Our Analysis shows that in case of such interactive applications, as the memory usage grows, the response time increases in a significant fashion resulting in poor user interactivity. Further, in cases where the data size exceeds the limits imposed by the available physical memory, the entire data cannot be loaded in memory for performing any kind of operation.

This paper introduces a way of effectively managing large quantity of VOTable data (of the order of millions of data points) for the purpose of plotting and analysis. The approach is to pre-process the data to create intermediate data structures, which are stored on the disk. These intermediate data structures are then used for plotting and/or analysis by the application. This technique has been successfully used for VOMegaPlot, the Java based plotting tool developed under the VO-India initiative. The technique is fairly generic and can be easily extended to astronomical data formats other than VOTable.

Pre-processing operation divides the input data into a number of blocks. For plotting and/or analysis, individual blocks are loaded as and when required which leads to lesser memory consumption and greater degree of user interactivity. Storing data in this fashion also helps in handling subsets of data as only those data blocks containing the subset can be loaded and used which results in faster operations.

VOMegaPlot supports various types of plots like scatter plot, density plot, projection plot and histogram. Following are the results of pre-processing and plotting operation for datasets, Tycho[+], Tycho-2[+] and UCAC2[*]. Results were obtained on a single CPU Pentium 4 (2.66 GHz) machine on Redhat Linux 9.0 platform.

	Tycho	Tycho-2	UCAC2
Data Size	1 million rows, 56 columns	2.5 million rows, 32 columns	48.3 million rows, 9 columns
Pre-processing time	18 minutes	30 minutes	3 hours and 26 minutes
Plotting time for scatter plot	9 seconds	22 seconds	5 minutes and 46 seconds

Notes:
[+]: Tycho and Tycho-2 datasets were pre-processed by keeping the memory size at 64 MB.
[*]: UCAC2 dataset was pre-processed with 256 MB memory size.