A usage summary of the VAST Challenge Datasets

Swetha Reddy, Manas Desai, Catherine Plaisant, *HCIL, University of Maryland*
Jean Scholtz, *PNNL*

Contact email *:* plaisant@cs.umd.edu

Technical Report

<19 October 2010>

Background

Visual Analytics research benefits from the availability of representative datasets and tasks to facilitate the development of tools and their evaluation. From 2006 to the present the VAST Challenge has contributed to making the datasets developed at PNNL known, and of course organized competitions around those datasets. This report summarizes the data we have gathered to document the use of the datasets and quantify the benefits to Challenge participants.

Datasets Made Available

- VAST 2006 Contest dataset (mostly text, news article)
- VAST 2007 Contest dataset (mostly text, news article)
- 4 VAST 2008 Challenge datasets (wiki edits, cell phone social network, spacio-temporal [boat migrations, and building evacuation])
- 3 VAST 2009 Challenge datasets (Badge and computer network traffic, Social network with geospatial component, video analysis)
- 3 VAST 2010 Challenge datasets (Text Reports, Spatio-temporal data, Genetic Sequences)

Challenge Participation

There were only six teams participating in the 1st (2006) VAST contest, then seven in 2007. The format of the contest was then changed in 2008 by introducing several mini challenges. Researchers could participate in more than one challenge. With the increase in number of datasets the number of entries also increased to 63. Out of the 63 there were six Grand Challenge (GC) entries and 67 mini-challenge entries. Twenty eight different organizations from thirteen countries submitted entries among which thirteen were student teams. For VAST Challenge 2009 there were 49 submissions among them five were for the Grand Challenge and the rest were for the mini challenges. The teams again came for 28 different organizations from 13 countries. In the following year for VAST Challenge 2010 there were 56 submissions among them five were for Grand challenge and rest for the mini challenges. The number of teams increased from twenty eight in 2009 to thirty three in 2010.

While the number of team remains limited, it is important to note that the number of participants represents a significant portion of the attendees of the VAST symposium.

Participation Table

Year	Total Entries	Number of Organizations	Student Teams	Participants
2006	6	6	3	~25
2007	7	7	3	38
2008	73	28	9	~88
2009	49	28	8	~105
2010	56	33	16	~153

Dataset downloads

The number of participants is not the true indicator of the use of datasets since the datasets are available after the contest as well. So we can use the number of downloads as a separate metric of usage. The number of downloads for each datasets is as follows.

Year	No. of downloads (as of October 2010)
2006	463
2007	189
2008	690
2009	905
2010	538

The high number of downloads of these datasets clearly suggest that many more people use the datasets than participants in the competition.

Comparing the email ids of all the users who downloaded these datasets, we could see returning users from 2006 through 2009. There are 65 users who used 2006 and 2007 datasets and similarly there are 69 users who used 2007 and 2008 datasets. Among these users there are **39** of them who download all three (2006, 2007 and 2008) datasets. Among the five hundred odd users of the 2009 dataset very few are users from previous years. These counts are not entirely reliable since they are based on the email ids that users enter while registering to download a datasets. A probable explanation for the few users returning in 2009 is that the topic change radically (from text to tabular data) some users might have had their colleagues and students download the datasets and might have shared it locally among themselves. We also know of few who have participated in the contest for all four years but have changed their email ids.

Use of the Benchmark Repository

With so many users using the datasets, it became necessary to have one constant location to hold information about various datasets, analysis, tools and research. Therefore a Visual Analytics Benchmark Repository was developed to bring together all the available benchmarks used for visual analysis and their uses.

The website lists the various benchmarks and their corresponding uses. It displays benchmark information like its description, location, solution etc. Users can view benchmarks based on their provenance and topics. It also allows users to add more benchmarks to the repository if they request a user id and password first. Benchmarks can be classified under various topics such as social network analysis, geo-spatial analysis etc. Users can add a brief description about the benchmark and specify its location.

The website also lets users to view the uses/analysis of the benchmarks. We initially populate the site with the competition entries but users can submit more uses to any of the existing benchmarks. Users can submit videos, webpage's or documents. They can also edit/update uses as they work on it we also list papers related to the benchmarks, provenances and visual analytics in general. Users can add papers and link them to appropriate entities. Overall the website is a collection of visual analytic benchmarks and related material for research, education and reference. The link to the website is http://www.cs.umd.edu/hcil/varepository.

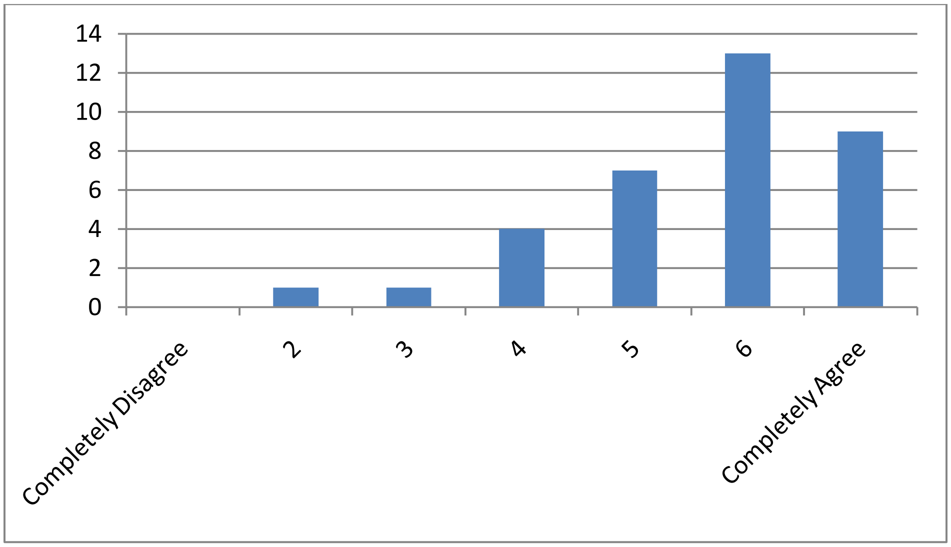
We started with all the VAST datasets and their uses (Challenge entries) and later also had few benchmarks from other research labs and universities. Finding papers that used the dataset was more difficult; we searched for papers which used any of the VAST dataset on the internet to add to the repository. Our search terms ranged from simple words like 'VAST challenge' and 'VAST contest' to specific names of imaginary people and places in the datasets like 'Isla Del Sueño', 'Alderwood' and 'Blue Iguanodon' etc. In our quest to add to the repository we informed all the VAST users through email about the repository and asked them to add any papers they might have written using the VAST dataset. The last update was done in 2009.

The repository currently holds 21 benchmarks, 175 uses and 44 publications. The website went live in September 2009 and had **2401** hits as on July 2010. We use Google Analytics to track the usage of the site and so far it has 1428 absolute unique visitors who spend an average of 4:30mins viewing the site. This short duration is appropriate since the site acts mostly as a portal to the other sites.

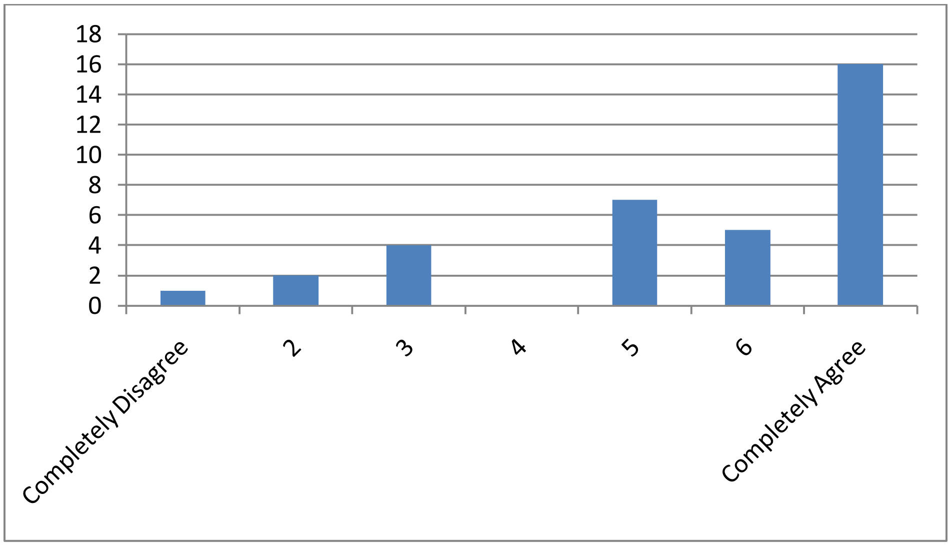
Results from Participant Survey

We conducted a survey of the participants of VAST Challenge 2009 with regards to the usefulness of the Challenge and the datasets. The survey was administered via surveymonkey, and an email was sent to the 2009 team in February 2010, i.e. 4 months after the symposium (so they had time to reflect on the long term benefit of the experience).

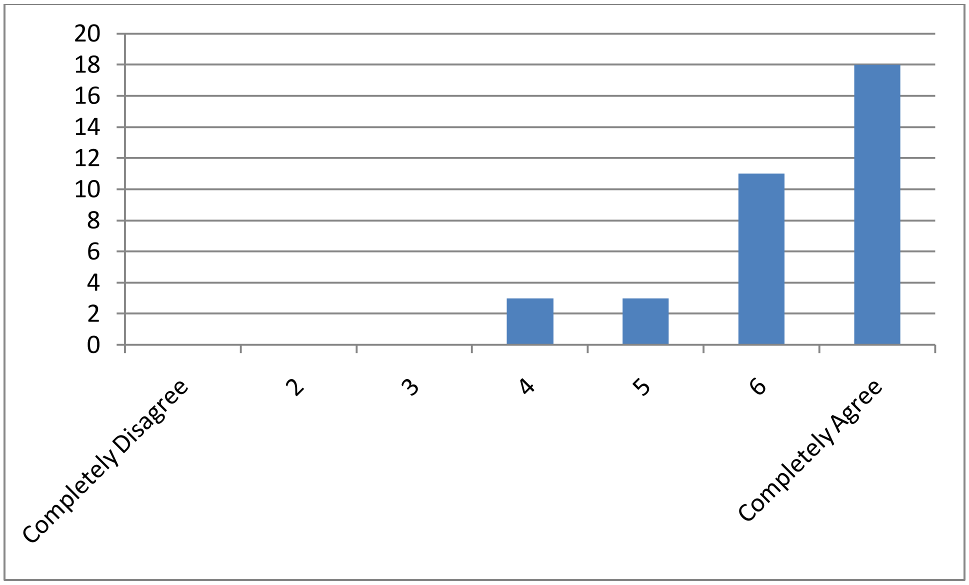
The survey had in all **35** responders. The results show that around 80% of the users agree that VAST Challenge allowed them to better understand the tasks that would be performed in an analytic task.



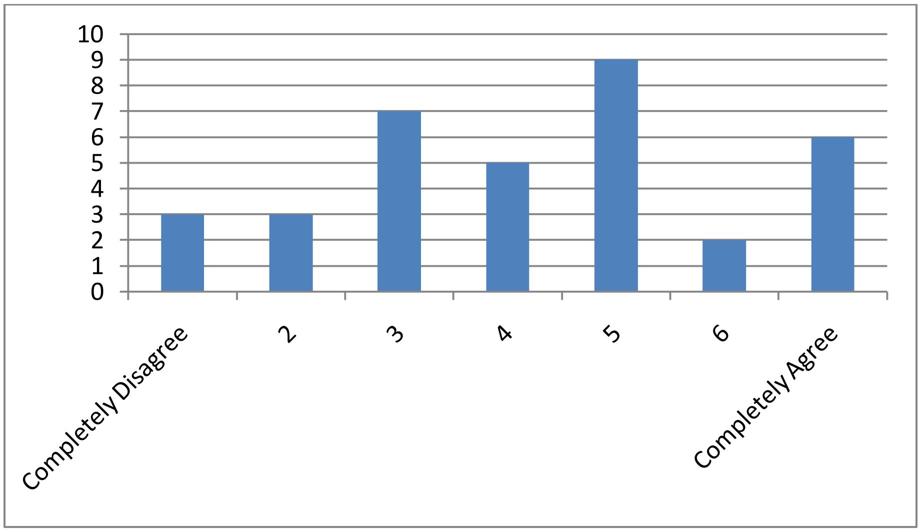
Most (around 80%) of the users agree that the Challenge helped them in improving their tools and sometimes even build new tools and approaches. One user even commented that they developed a novel system which would not exist if not for the VAST challenge.



The majority also thought having ground truth in the dataset was important and one of the user commented that it increases the sense of puzzle solving.



Another noteworthy result was that the users were quite divided on the usefulness of the reviews on the submissions given by the reviewers and analysts.



Overall the survey results are very satisfactory and support the effort undertaken by the VAST Challenge committee.

We even collected feedback from users through email and found out that the datasets are used to teach in class at SFU Business School. Few users also used the datasets for internal testing of their information visualization tool.

Clearly the datasets have been used by number of people for various purposes.

Conclusions

Ever since the introduction of the VAST Challenge in 2006 it has been growing in strength with each year. The number of organizations, the total number of participants and the number of student participants has increased significantly. As mentioned above, the datasets have been useful in various activities and we can anticipate that in future VAST Challenge will play much more important and vital role in the field of information visualization.

To learn more about the VAST Challenge

Grinstein, G., Konecni, S., Plaisant, C., Scholtz, J., Whiting, M., VAST 2010 Challenge: Arms Dealings and Pandemics, Proc. of VAST 2010 Conference, IEEE (2010) 267-268

Scholtz Jean, Developing Qualitative Metrics for Visual Analytic Environments. In the *Proceedings of BELIV '10 (Atlanta, USA).*1-7

Whiting Mark, Generating a Synthetic Video Dataset. In the *Proceedings of BELIV '10 (Atlanta, USA)*. 43-48

Grinstein, G., Plaisant, C., Scholtz, J. and Whiting, M., VAST 2009 Challenge: An Insider Threat, *Proc. of IEEE VAST. 2009* 243 - 244

Whiting, M., North, C., Endert, A., Scholtz, J., Haack, J., Varley, C., Thomas, J., VAST Contest Dataset Use in Education, *Proc. of VAST 2009 Symposium*. 115 - 122

Costello, L., Grinstein, G., Plaisant, C. and Scholtz, J., Advancing User-Centered Evaluation of Visual Analytic Environments through Contests,*Information Visualization*, 8 (2009) 230–238).

Grinstein, G., Plaisant, C., Laskowski, S., O'Connell, T., Scholtz, J., Whiting, M., VAST 2008 Challenge: Introducing Mini-Challenges, *Proc. of IEEE Symposium on Visual Analytics Science and Technology* (2008) 195-196.

Plaisant, C., Grinstein, G., Scholtz, J., Whiting, M., O'Connell, T., Laskowski, S., Chien, L., Tat, A., Wright, W., Gorg, C., Liu, Z., Parekh, N., Singhal, K., Stasko, J. *IEEE Computer Graphics and Applications* 28, 2, March-April 2008, pp.12-21 (2008)

Whiting, M., Haack, J., Varley, C., Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software, *Proc. of BELIV'08, BEyond time and errors: novel evaLuation methods for Information Visualization*, ACM (2008).

Plaisant, C., Fekete, J. D., Grinstein, G., Promoting Insight Based Evaluation of Visualizations: From Contest to Benchmark Repository, *IEEE Transactions on Visualization and Computer Graphics*, 14, 1 (2008) 120-134

Grinstein, G.; Plaisant, C.; Laskowski, S.; O'Connell, T.; Scholtz, J.; Whiting, M.; VAST 2007 Contest - Blue Iguanodon, *Proc. of the IEEE Symposium on Visual Analytics Science and Technology*, VAST 2007, pp 231 – 232.

(See also http://www.cs.umd.edu/hcil/semvast/)