Visualizing Large Datasets of Images in Web Analytics

Dhawal Mujumdar
University of California, Berkeley
102 South Hall, Berkeley CA
dhawal@ischool.berkeley.edu
510-415-5563

ABSTRACT
Web analytics tools fail to report information about rich media elements like images and videos. The usage of rich media elements on websites is increasing exponentially over the time. In this paper, we introduce new visualization technique to visualize rich media elements as a part of web analytics framework. We’ve used data from recent user studies conducted on our image intensive mobile application. Microsoft Pivot Viewer was used to create visualizations. We discuss salient features of our visualization technique, its advantages and shortcomings. We also discuss the scope for future research work.

Author Keywords
Web analytics, data visualization, rich media elements.

ACM Classification Keywords
H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems s.

INTRODUCTION
Web analytics is a technique of collecting and reporting Internet usage statistics and data. Web analytics is used for multiple purposes like measuring traffic, user research, marketing campaigns, advertising spend etc. Web analytics tools are particularly important in understanding how users engage with your website. Web analytics has matured over the last decade and remains the cornerstone of continuously evolving World Wide Web.

Given the amount of data that is collected using web analytics, every solution for web analytics (ex. Google Analytics) provides number standard data visualizations like

- Demographics and System Statistics
- Internal Search Information
- Visitor Paths
- Visitor Types

Unfortunately, no attempts have been made to analyze information about rich media elements like images and videos. We are living in the world where the proliferation of rich media elements is happening and it’s necessary to include visualizations to succinctly convey the message about how users are interacting with these elements.

To solve the aforementioned problem, we’ve developed an interactive visualization using Microsoft Pivot Viewer [3]. Pivot viewer is a tool designed to interact with massive amounts of data in ways that are powerful, informative, and fun. We will be discussing the intent to use this tool and the techniques used in the implementation section of this paper.

RELATED WORK
Existing research literature on the web analytics in the context of rich media elements is limited; however sufficient research has been done over visualizing large datasets of images in the context of image retrieval.

Rodden et al. investigated how people manage and organize their image collections. Many of these techniques involved in management include use metadata, such as keywords, time, as a basis of photo organization.

Naaman et al. [2] developed a technique to cluster images based on the time and location metadata. Chen et al. [1] devised the method for similarity based image browsing. They organized and clustered the images based on the color histograms and textures. Moghaddam et al. [4] introduced a method to visualize images based on their mutual similarities visualized on a 2-D screen. The proposed technique shows a more perceptually intuitive and informative visualization for images.

Most of the existing research covers qualitative aspects like color, size, shape, texture etc. in visualizing images. However, it fails to address quantifiable elements, which are important in web analytics.

DATASET
We are currently doing a research on image-recognition CAPTCHA (see fig 1.). For this research project we had
built an image database of ~150 images. We were conducting detailed user studies with end-users that were using these CAPTCHAs.

Fig 1: Image recognition CAPTCHA that was used to build dataset

We recorded every interaction of end-users with CAPTCHAs. This data consisted of image categories, image type, image positions, image clicks etc. Recorded data served as the dataset for creating our visualizations.

PURPOSE OF VISUALIZING GIVEN DATASET

Every data visualization has a purpose; the purpose of visualizing this data is as follows:

1. Quick snapshot of all the images that are served using CAPTCHAs so far
2. To monitor all the images that users can easily associate with their respective categories
3. Collect relevant metrics

We will be discussing how our visualization addresses aforementioned goals in the discussion section.

VISUALIZATION TOOL

We chose Microsoft Pivot Viewer to create visualizations using our datasets. Pivot Viewer uses Microsoft Silverlight to display visualizations and it works on Microsoft .Net platform. Pivot Viewer contextualizes information in much more natural way for humans to digest large amounts of information without losing their way. It also uses faceted navigation to present end-results. Faceted classification allows multiple assignments to any images, which is important in large datasets of images.

Searching large data using Pivot Viewer is comparatively easy. It uses deep-zoom technique, which makes it easier to drill down to specific data. Using Pivot Viewer it is easy to perform data analysis. For every image associated metadata can be viewed easily (fig 4).

IMPLEMENTATION

As discussed earlier, we were using image corpus from our user studies on image recognition CAPTCHAs. Out of ~150 images, ~60 images were served in user studies by the time we collected this data. We logged all the user interactions (like image clicks) in XML files. Of all the data we collected, we chose image category, image type (target: image that should be chosen to clear CAPTCHA challenge, random: image that is irrelevant), image position (in the grid 3X3) and Image Clicked (yes or no). Using this data we were able to generate following visualizations: Fig 2&3, presents overall visualization that presents dataset. Fig 4 shows how data can be filtered and drilled down to specific details. Fig 5, shows metadata associated with each individual image.

Fig 2 & 3: Visualizing all the images

Fig 4: Filtering information using facets
DISCUSSION
Using this visualization technique, users can quickly visualize all the images that are served using CAPTCHAs. Users can do thorough data analysis on the images that are difficult to choose and the ones that are easy to pick. For example, user can see the target images that are not consistently selected by end-users (in order to pass CAPTCHA). Similar analysis can be done on image categories i.e. the categories that contain highly representative images that are successfully selected by end-users. We believe the technique sufficiently addresses all the goals discussed in the previous sections. However, to support this claim we need conduct thorough user studies, which are discussed in the “Future Work” section.

We also discovered that our visualization technique has few limitations. We wanted to address the problem of visualizing any dataset in web analytics. However, this technique addresses application specific (image recognition CAPTCHAs) goals. After using Pivot Viewer we’ve realized that this tool is best to use with really large datasets of images. Thus, if we are going to develop similar technique as a part of general web analytics tools (like Google Analytics) then it will not be useful for users that use few images on their websites. Lastly, we discovered that Pivot Viewer itself has many limitations. Pivot Viewer is not an ideal tool to do qualitative analysis over the images. It does not provide any insights on the image like hue, size, shape, texture etc. and thus these vital details are left out of comparisons using Pivot Viewer. Users can add this information as facets but it is not provided by default.

FUTURE WORK
We were not able to conduct any user studies for our visualization technique. This remains the highest priority for our future work. Conducting user studies will give us insights into how users are using this product and evaluate overall effectiveness of our product.

In future we want to extend this visualization technique that should work on any dataset of images as a standalone tool or as a part of existing web analytics frameworks. Given the limitations of Microsoft Pivot Viewer, we want to create an open source visualization library in advanced technologies like HTML5 and jQuery. In the said library we want to include a feature that can do more qualitative comparisons of the images.

SUMMARY AND CONCLUSION
We have presented a new visualization technique to visualize rich media elements like images using web analytics framework. This technique can visualize large datasets of images efficiently and assists users with data explorations. We believe this is an early research work that can be extended beyond existing dataset. We believe this work can lead to promising directions in visualizing rich media elements.

REFERENCES