Approximate Algorithmic Image Matching to Reduce Online Storage Overhead of User Submitted Images

Braden D. Licastro (Professor Robert S. Roos, Project Advisor)
Department of Computer Science, Allegheny College
Major: Computer Science, Minor: Economics

Image Sharing Websites

Basic Idea: Image sharing websites allow the public to upload an image of their choosing to the internet. This process lends itself to duplicate data. Below are examples of the more popular image sharing sites.

flickr photobucket Instagram imgur

Figure 1: Image sharing website examples.

Motivation

- As of May 2013, nearly 500 million images were shared each day. This is expected to double by May 2014. [1]
- Approximately 20% of this data is estimated to be duplicate. [2]
- By eliminating this duplicate data, companies can save roughly $1.8 million annually.

Cost Reduction Techniques

Currently several technologies are used to reduce the costs associated with storing the shared data including:

- File size restrictions
- Per-user restrictions
- Subscription services
- File upload compression
- Upload expiration times

Duplicate Detection Algorithm

To further reduce the costs of hosting user submitted image data, a two-stage algorithm was created allowing for accurate detection of duplicate images.

Stage I: Exact file detection

- Detects identical files using unique MD5 hashes, a unique text representation of a file. If the hash matches, an exact duplicate has been found as seen in Figure 2.

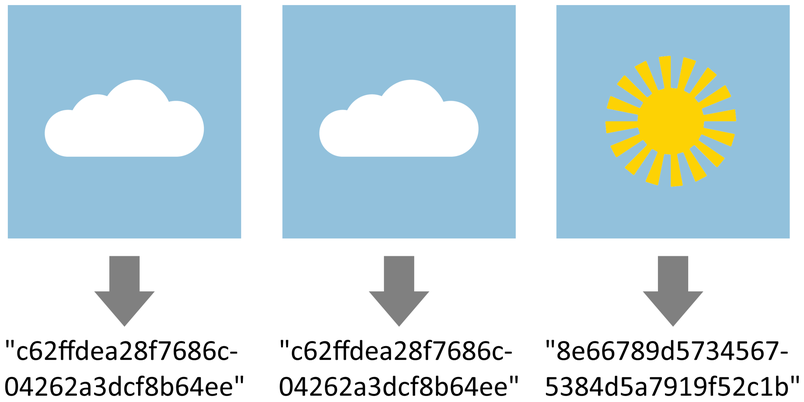

Figure 2: Images and their associated hashes.

Stage II: Partial match detection

Color profile matching:

- Detects identical files by determining the average number of times each color occurs, and creates an MD5 hash from the resulting counts.
- If color profiles have differing hashes, the images are unique, otherwise further analysis is needed.

Pixel-by-Pixel Analysis:

- If color profiles match, images are resized to a set width and height. Individual pixel colors are analyzed one by one, and if the two images differ by more than a set threshold, they are unique, otherwise duplicates.

Evaluation Metrics and Results

When evaluating the efficacy of the research, several groups of images were developed containing unique and duplicate images in various sizes. Metrics recorded include processing time, detection rates, and storage requirements.

Processing time:
Processing times increased, but remained less than two seconds in observed scenarios, as seen in Figure 3.

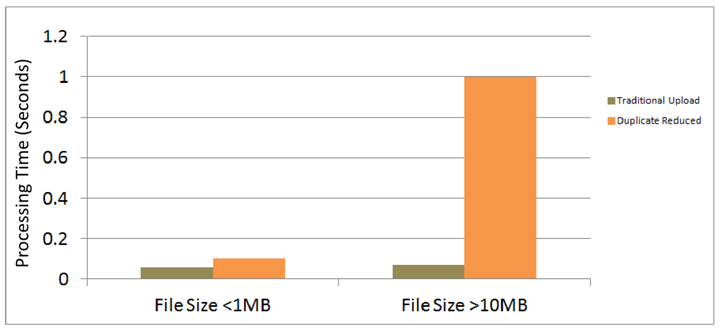

Figure 3: Processing time for associated upload methods.

Detection rates:
Photographic image duplicate detection correctly identified images more than 85% of the time, and in nearly every case for low-detail computer generated patterns.

Storage requirements:
Test data containing an average of 20% duplicate data was reduced by 12% once processed using the detection algorithm. This equates to $1.08 million in savings annually.

References

[1] All Things Digital. Meeker: 500 Million Photos Shared Per Day and That's on Track to Double in 12 Months. http://goo.gl/Ht77pV

[2] NTP Software: Survey Says Nearly Two-Thirds of Files on Primary Storage are Stale. http://goo.gl/udiY47

Contact Author: Braden D. Licastro <licastb@allegheny.edu>