

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

Braden D. Licastro

Allegheny College, USA

licastb@allegheny.edu

April 18, 2014



ALLEGHENY COLLEGE

Photo Sharing Service

Definition

Photo sharing is the publishing or transfer of a user's digital photos online, thus enabling the user to share them with others [1].

Example Services

The logo for imgur, featuring the word "imgur" in a bold, black, sans-serif font with a small green dot above the "i".The logo for shutterfly, featuring the word "shutterfly" in a playful, rounded font with a small cluster of colorful dots above the "s". Below the word is the tagline "where your pictures live" in a smaller, sans-serif font.The logo for flickr, featuring the word "flickr" in a bold, sans-serif font with "flick" in blue and "r" in pink. A small trademark symbol is visible.

Remembering the Facts

- 500 Million images shared daily [2]
- Daily image shares expected to double in 2014 [2]
- Approximately 20% of stored data is duplicate [3]
- Eliminating duplicates can save roughly \$1.8 million annually at current sharing levels¹

¹ Assuming 2013 averages of \$.05 per gigabyte and 1MB image size[3].

Goals

Website Creation

Created a flexible website framework that was able to imitate an image sharing service.

The Algorithm

Employed a series of checks and algorithms to find and eliminate duplicate and near-duplicate images at the time of upload.

Result Compilation

Website generates real time directory file count, directory size, and collects time taken per image upload.

Website Details

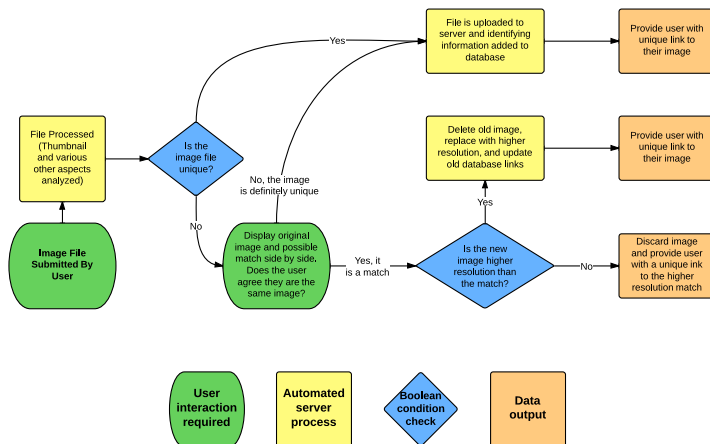


Figure: Duplicate Identification Process

Website Details

Database: thesisDB Table: shareTracker			
Name:	Type:	Description:	Extra
ID	int(11)	Gives every image a unique ID	Primary Key Auto-Increment
lLookup	varchar(6)	Unique URL ID Lookup	Not Null
lName	varchar(21)	Images file name on server	Not Null
directory	varchar(15)	File location from server root	Not Null
uMethod	int(1)	Upload method used	Not Null
hash	varchar(40)	Hash of the image for exact dup matching	
fingerprint	varchar(32)	The MD5 fingerprint of the histogram array.	
processTime	int(11)	Upload time, from start to completion	Not Null

Figure: Schema for File Uploads

Generation of Comparison Data

- Every upload added to baseline directory
- Only unique images added in duplicate-reduced directory
- Time taken for each upload is recorded in the database
- Real time directory statistics updated on every page load

Test Cases

All cases are composed of half computer-generated, half photographic imagery.

I : Small Images $< 1MB$; No Duplicates

II : Small Images $< 1MB$; 20% Duplicates

III : Large Images $> 1MB$; No Duplicates

IV : Large Images $> 1MB$; 20% Duplicates

Performance Calculation

Using this equation, processing time, directory counts, and directory size were calculated.

$$\left(\frac{Base - Reduced}{Base} \right) * 100 = \%ImprovementOverBase$$

Figure: Percent efficiency over base case.

Image Matching Demonstration

View the live website...

► Demo Login

References



[Dictionary.com.](#)

Definition of Photo Sharing.

www.dictionary.com, 2013.

[Online; accessed 20-November-2013].



[All Things Digital.](#)

Meeker: 500 Million Photos Shared Per Day and That's on Track to Double in 12 Months.

[http://allthingsd.com/20130529/](http://allthingsd.com/20130529/meeker-500-million-photos-shared-per-day-and-thats-on-track-to-double-in-12-months/)

[meeker-500-million-photos-shared-per-day-and-thats-on-track-to-double-in-12-months/](http://allthingsd.com/20130529/meeker-500-million-photos-shared-per-day-and-thats-on-track-to-double-in-12-months/), 2013.

[Online; accessed 20-November-2013].



[NTP Software.](#)

Survey Says Nearly Two-Thirds of Files on Primary Storage Are Stale.

[www.ntpsoftware.com/pressroom/](http://www.ntpsoftware.com/pressroom/survey-says-nearly-two-thirds-files-primary-storage-are-stale)

[survey-says-nearly-two-thirds-files-primary-storage-are-stale](http://www.ntpsoftware.com/pressroom/survey-says-nearly-two-thirds-files-primary-storage-are-stale), 2013.

[Online; accessed 31-October-2013].

Questions? Comments?