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

imgur

shutterfly
where your pictures live

flickr

Instagram

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 levels[1]

[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

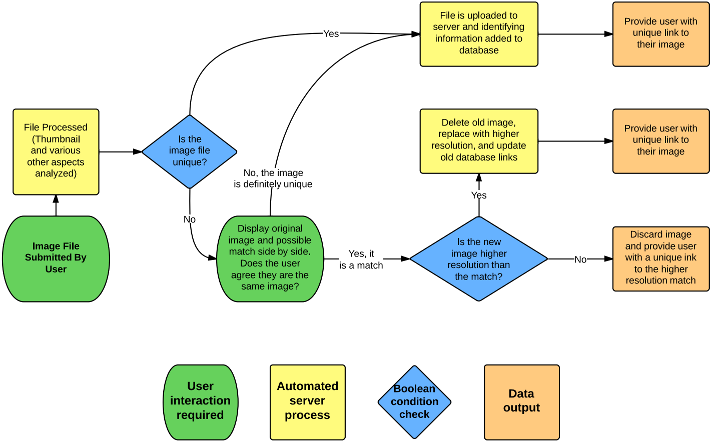
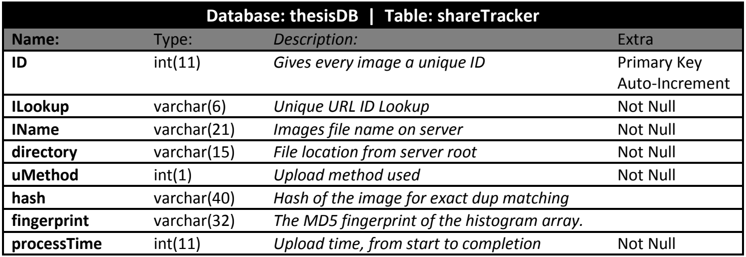


Image File Submitted By User
File Processed (Thumbnail and various other aspects analyzed)
Is the image file unique?
Yes
No
No, the image is definitely unique
File is uploaded to server and identifying information added to database
Provide user with unique link to their image
Display original image and possible match side by side. Does the user agree they are the same image?
Yes, it is a match
Is the new image higher resolution than the match?
Yes
Delete old image, replace with higher resolution, and update old database links
Provide user with unique link to their image
No
Discard image and provide user with a unique ink to the higher resolution match
User interaction required
Automated server process
Boolean condition check
Data output

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
ILookup	varchar(6)	*Unique URL ID Lookup*	Not Null
IName	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

- Case I: Small Images $< 1MB$; No Duplicates
- Case II: Small Images $< 1MB$; 20% Duplicates
- Case III: Large Images $> 1MB$; No Duplicates
- Case 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/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/survey-says-nearly-two-thirds-files-primary-storage-are-stale, 2013.
[Online; accessed 31-October-2013].

Questions? Comments?