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

Braden D. Licastro

Allegheny College, USA

licastb@allegheny.edu

November 25, 2013

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

Motivating 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 $18 million annually at current sharing levels[1]

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

Goals

Website Creation

Create a flexible website framework that is able to imitate an image sharing service.

The Algorithm

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

Result Compilation

Website must generate real time directory file count, directory size, and collect time taken per image upload.

Website Details

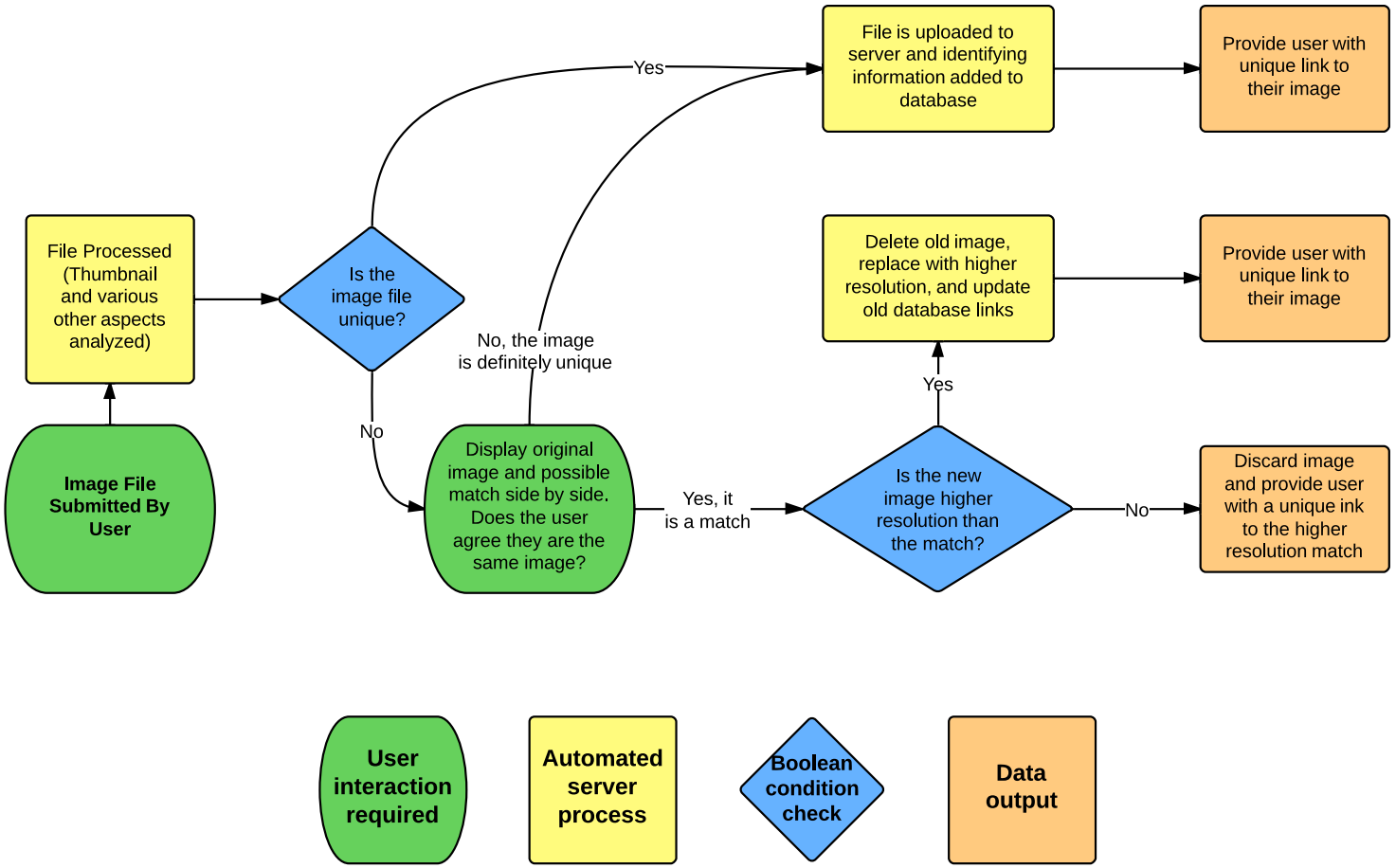


Figure : Duplicate Identification Process

Website Details

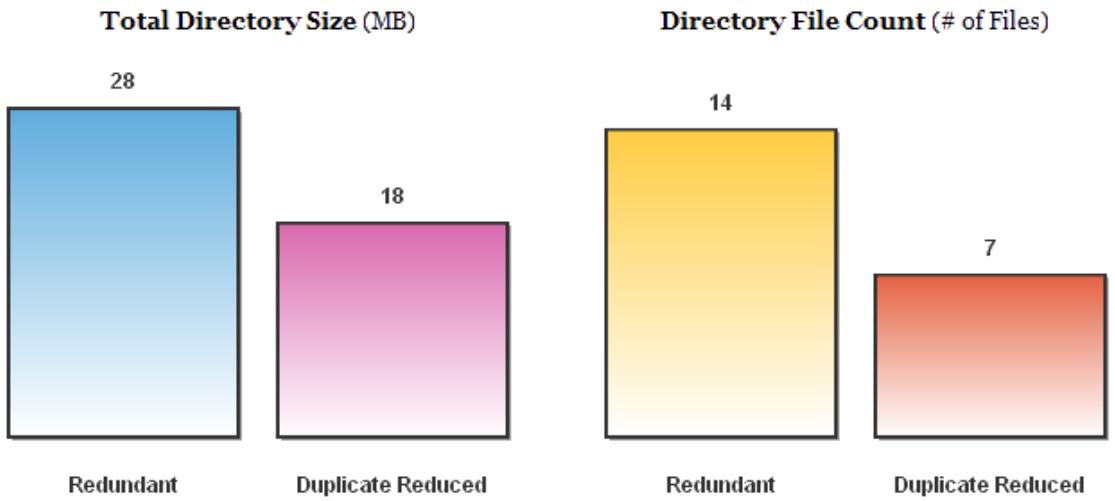
Column	Type	Comments
ID	int(11)	Gives every image a unique ID
ILookup	varchar(6)	Unique URL ID Lookup
IName	varchar(17)	Images file name
directory	varchar(12)	File location from virtual server root
uMethod	int(11)	Upload method
histogram	varchar(32)	Hashed histogram of dup-reduced images
processTime	int(11)	Upload time to completion

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

Real-Time Statistics



VITAL STATISTICS

STORAGE UTILIZATION

Traditional Upload Directory:
28.84 MB
Duplicate Reduced Directory:
18.29 MB

DIRETORY FILE COUNT

Traditional Upload Count:
14 Files
Duplicate Reduced Count:
7 Files

Figure : Schema for File Uploads

Performance Calculation

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

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?