

CrowdFlower

Customer Report

Business Listing Verification and Collection

CrowdFlower, Inc.
2111 Mission St.
San Francisco, CA 94110

Solutions@CrowdFlower.com
(415) 621-2343
www.crowdfLOWER.com

Overview

This report describes the verification and correction of 99,987 business listings that client provided to CrowdFlower (Phase I). Additionally, CrowdFlower gathered enriched attributes (Phase II) on a sample of client's data. CrowdFlower was able to raise client data accuracy from ~70-80% accuracy to >97% accuracy for Phase I, and gather enriched attributes at ~97% accuracy.

Local Search

Local search companies all face the same problem. After buying data from the same data providers (Infogroup, Acxiom, etc.), then integrating and cleaning the data programmatically, they're left with the same ~75-85% accurate data all their competitors have. From there, local search companies seek to gain an edge via large investments in human capital for (1) cleaning of core business listing data (URL, name, phone number, and address) to 95%+ accuracy, (2) expanding their dataset of full businesses listings, and (3) enhancing listings with enriched data (e.g., hours of operation, pictures from business homepages, information on restaurant and hotel amenities, business closures, etc.).

CrowdFlower helps local search companies cost-effectively achieve these goals. CrowdFlower trains and continuously evaluates over one million workers who use custom-built web tools to verify and correct full listings, acquire URLs and other data for partial listings, and acquire enriched attributes at higher accuracy, lower cost, and with fewer painful headaches than traditional outsourced/internal solutions. Best of all, there's no up-front investment like that of hiring internal workers or training outsourced workers; CrowdFlower's workforce can be turned on and off on demand and companies pay only for the work they receive.

Phase I Project Summary

Due to the high proportion of medical professionals in the dataset client provided CrowdFlower, CrowdFlower split client's dataset into a medical professionals dataset and a standard business set using a combination of the client-supplied category data and CrowdFlower workers. Name, URL, Address, and Telephone Number elements were collected online from official websites of the businesses in question. For businesses, if such a website was not discovered, CrowdFlower workers determined if the business was closed or not by searching Yelp.

A unit is one business or one Medical Professional. Each unit was routed through an integrated workflow of crowdsourced workers working through custom-designed interfaces. At each step in the workflow, several workers' judgments were algorithmically aggregated to one trusted answer based on workers' individual accuracy. Individual worker accuracy is assessed by test "gold" units that workers complete in the course of judging client data.

Accuracy

An audit was performed on 300 randomly selected units from both the business and the Medical Professionals sets. For each set, CrowdFlower audited 200 units with client-provided candidates, as well as 100 units with no client-provided URL for which CrowdFlower found a new URL.

Businesses

Precision:

Businesses - Verified Candidates				
	URL	Name	Phone	Address
CrowdFlower Precision	99%	100%	99%	99%
Client Precision	71%	79%	95%	78%

Businesses - Found Candidates				
	URL	Name	Phone	Address
CrowdFlower Precision	99%	100%	98%	99%
Client Precision	NA	71%	99%	84%

Recall:

Recall defined: "if there was a relevant record to return, what % of the time did CrowdFlower return it?" Of the units provided to CrowdFlower by client, CrowdFlower returned units at 92% recall.

Medical Professionals

Precision:

Medical Professionals - Verified Candidates				
	URL	Name	Phone	Address
CrowdFlower Precision	98%	98%	98%	98%
Client Precision	68%	NA	89%	83%

Medical Professionals - Found Candidates				
	URL	Name	Phone	Address
CrowdFlower Precision	97%	99%	94%	97%
Client Precision	NA	NA	81%	84%

Recall:

Of the units provided to CrowdFlower by client, CrowdFlower returned units at 91% recall.

URL Verification and Acquisition

Source data without a URL candidate or that had its candidate URL rejected by workers had their URLs searched for with custom search engines built on top of Google and DuckDuckGo. After URLs were discovered, another set of workers verified that the discovered candidates were correct.

The remainder of the listings, for which a primary website was not verified or discovered, workers indicated if they could positively identify if the business in question was closed through a search of Yelp.

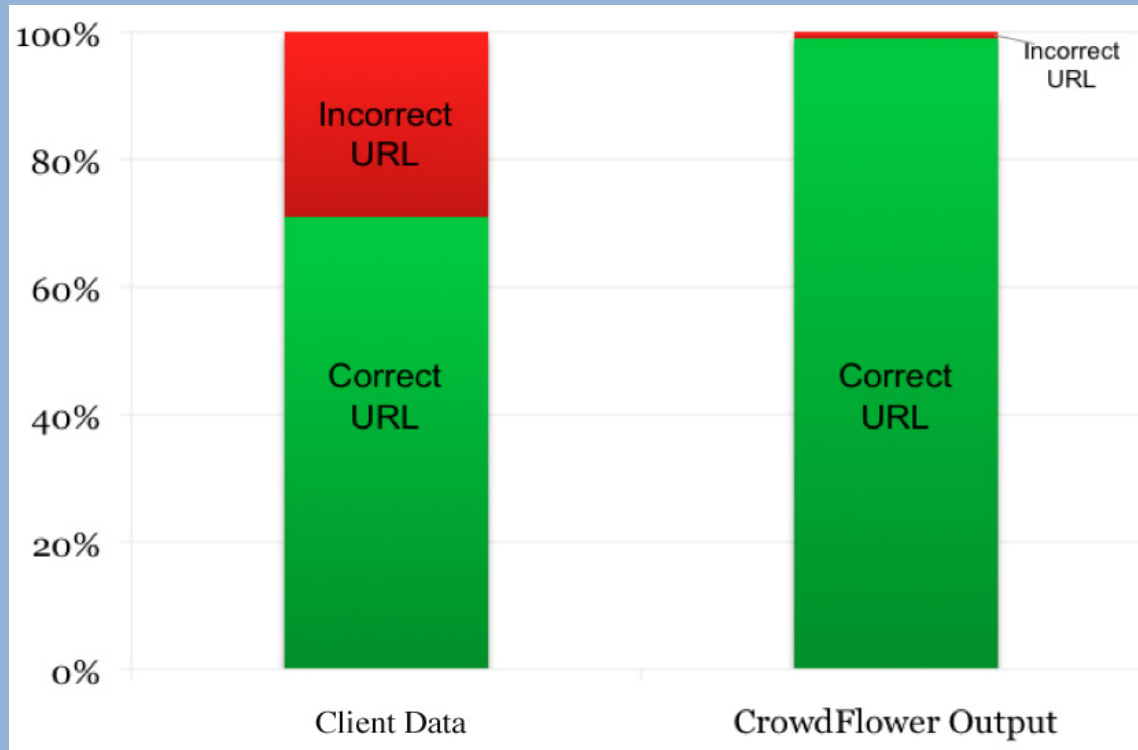
CrowdFlower's results are presented in the two tables below.

Business		
	URL	CF
Correct URLs	17,135	22,828
Total Units	51,558	51,558
URLS	24,134	23,059
% Correct	71%	99%

Medical Professionals		
	Client	CF
Total Units	48,429	48,429
URLS	5,618	14,758
Correct URLs	3,820	14,392
% Correct	68%	98%

For medical professionals, CrowdFlower returned doctor-specific pages where available

URL Accuracy



Name

For businesses, source units that had a URL verified or acquired then had their source unit Name element verified and/or cleaned. Names were considered correct only if they exactly matched one form of the name presented in the logo or text of the business's website (e.g., identical capitalization, identical punctuation). If the name was found to be incorrect, workers input the correct name from the website text.

For medical professionals, CrowdFlower instructed workers to return both the name of the medical professional and the name of the medical professional's practice, in order to have the data uniformly formatted.

Source data with a Name element that closely matched the name (or a common misspelling) of one of the top several hundred chains in the USA had the correct name appended programmatically.

Phone

For both medical professionals and businesses, source units that had a verified or acquired URL had their source unit Telephone Number element verified or collected. CrowdFlower instructed workers to verify if the source data was correct, and if not, paste in the correct phone number for that business's/medical professional's location (if available).

Address

For both medical professionals and businesses, source units that had a verified or acquired URL then had their source unit Address element verified. If the address was not verified as correct, workers would fix each address element so that it was correct. CrowdFlower measured address accuracy based on the complete address (address line, city, state, and zip).

Phase II Project Summary

In addition to the Name, Address, Phone, and URL, CrowdFlower also collected additional attributes. Per discussions with client, CrowdFlower agreed to test collecting data for business Hours of Operations, Restaurant Amenities, Hotel Amenities, as well as assess the accuracy of client's category data. The results are discussed below.

Hours of Operation

CrowdFlower tested collecting Hours of Operation. CrowdFlower excluded some businesses based on category, such as Philanthropy and Colleges & Universities, where it was obvious that Hours of Operation would not be available or meaningless.

CrowdFlower's audit indicated that its accuracy for this sample was 97%.

Negative Categorization

The negative categorization task was aimed at determining whether or not the source categories provided to CrowdFlower by client were correct. Workers visit a website for the business and determined which client categories were incorrect.

Based on CrowdFlower's internal audits, client's category accuracy to begin with was 92%, and CrowdFlower was able to increase that accuracy to 99%.

Restaurant Menus

CrowdFlower conducted a test on restaurants designed to acquire the URLs for either a specific menu or the section of the website where menus are found.

Based on CrowdFlower's internal audit, workers achieved a 99% precision score and a 96% recall score on finding websites with menus.

Hotel Amenities

CrowdFlower instructed workers to collect whether source data hotels offered the most common hotel amenities. These included:

- High Speed Internet / WiFi
- Business Center
- Fitness Center
- Pool
- Onsite Restaurant or Bar
- Complimentary Breakfast
- Pet Friendly

Based on CrowdFlower's internal audit, workers were found to have a precision score of 99% and a recall score of 98%.

Restaurant Amenities

CrowdFlower instructed workers to collect whether source data restaurants offered the most common restaurant amenities.

These included:

- Delivery
- Catering
- Online Ordering
- Accepts Reservations

Based on CrowdFlower's internal audit, these attributes were collected with a precision score of 95% and a recall of 88%.

For more information contact:

Solutions@CrowdFlower.com

415 – 651 – 4485

www.crowdfLOWER.com