INFORMATION EVOLUTION
Reliable, Experienced, Affordable Human Bandwidth.

CASE STUDY

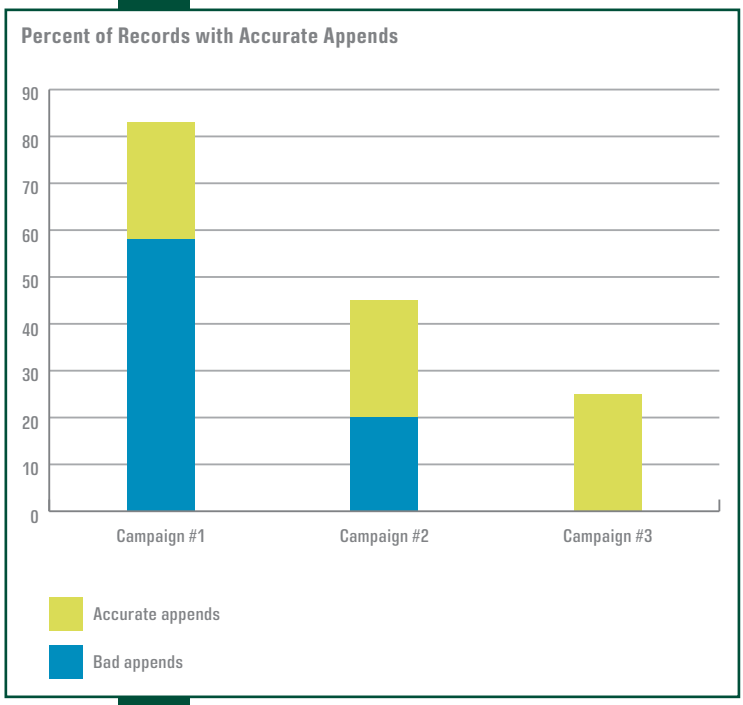
Managed Crowdsourcing

Effective Crowdsourced Data Appending

February 2011

Overview

Information Evolution's first crowdsourcing project involved appending URLs and email addresses to a database of Italian company names. IEI ran three different crowdsourcing campaigns for the job, each using a different methodology. We first tried posting the project on the Amazon Mechanical Turk marketplace. For the next campaign, we tweaked the process to address issues we experienced the first time around. Finally, we decided to add more software to the mix. Comparing these processes and their results provides some insight into managed crowdsourcing best practices.



The first campaign yielded a high append rate, but most of the appends were incorrect. The second campaign increased the percentage of accurate appends and lowered the overall append rate at the same time. The majority of appends were correct, but there was still a large minority of inaccurate appends. The third campaign yielded an even lower append rate, and, most importantly, completely eliminated inaccurate appends.

Process

The process for each successive campaign became more complex. The first (Campaign #1) was simply run through Amazon Mechanical Turk without any added checks and balances. The second (Campaign #2) also ran through Mechanical Turk, though this time the general task instructions were rewritten and clarified based on our experiences with and the results of Campaign #1.

On the third try (Campaign #3), instructions were rewritten again and some automated processes were added. The project used CrowdControl software, which runs on top of the Mechanical Turk API. URL and email validation ensured that accepted results were properly formatted URLs (www.companyname.com) and email addresses (XYZ@co.it). Invalid URLs, for example names such as Applegate, Europages, Facebook, OneSource, YouTube, Hoovers, DNB, PagineGialle, and yell.co.uk, were rejected automatically. In addition, URL "pinging" meant that only active, functioning websites were accepted.

Results

Metric	Campaign #1	Campaign #2	Campaign #3
URL append rate	83%	45%	25%
Email append rate	64%	50%	25%
Avg. time per assignment	1:13	1:33	1:33
Effective hourly rate per worker	$0.98	$0.77	$1.55

Initial submission quality was very poor for Campaign #1, a little better for #2, and 100 percent accurate after the IEI QC check for Campaign #3. Though there was improvement from Campaign #1 to Campaign #2, both required days of QC time. QC on Campaign #3, in contrast, took only a few hours. Overall quality of the final deliverable improved substantially over the course of the three campaigns.

Analysis

An examination of Campaign #1 data showed three types of false positive returns. The most common were sites such as YouTube, Facebook, OneSource, Coop Biz, EuroPages, and several others entered as company URLs. Second,

an alphabetical review of URLs showed that different companies with similar names sometimes had the identical URL listed. Third, some URLs that were accepted were just the full company name placed between "www" and "com," and did not lead to live websites. Beyond the clearly false returns, in several instances Campaign #1 allowed parent company web addresses as acceptable. Campaign #1 allowed the collection of a secondary email addresses which caused inconsistent data gathering. Specifically, in some cases a worker would find two email addresses and enter both without considering which should really be the primary address.

During Campaign #2, all false positive web sites were removed using find and replace processes. We used a "triple-check" formula in Excel to highlight additional inaccuracies, which caused even more URLs to be identified as false and discarded. Additional review processes ensured that duplicate URLs—cases where when the same URL was used for two or more different companies—were eliminated. Finally, Campaign #2 discarded parent company URLs and listed NIF (no information found) when a subsidiary did not have its own independent web page.

Campaign #2 netted fewer URL appends because of more rigorous data vetting developed using lessons learned from Campaign #1. It produced a higher percentage of accurate appends along with a 20 percent lower URL return rate as compared to Campaign #1 because of a reduction in "false positives."

Campaign #2 produced a 5 percent increase in appended emails addresses over Campaign #1. Campaign #2 also turned up a number of instances in which a subsidiary company with no independent website did in fact have a direct email address listed at a parent company website. As previously mentioned, Campaign #2 did not accept parent company URLs. Subsidiary email information found within parent company websites was, however, included. For Campaign #1, in contrast, if a company URL was returned as "NIF," the worker would automatically put "NIF" in the company email.

Campaign #3 produced lower URL and email append rates and cost more to run, but data quality was high enough to justify the expense. Campaign #3 produced a fully automated URL and append rate of 25 percent with no need for manual confirmation or rigorous QC by IEI. A quick QC confirmed a 100 percent accuracy rate.

The amount paid to crowdsourced workers grew while the append percentage declined because of the near total absence of false positives. While the append rate decreased, the accuracy of the appended data increased dramatically.

Generally, crowdsouce workers seem to prefer to enter some data, even if it is incorrect, rather than use "NIF" because they fear they will not get paid if they can't come up with an answer. In fact, of course, the opposite is true. If

a company has no web site, then that's the relevant information and an "NIF" is the answer that pays. The project setup must make it very clear to turkers that "NIF" is a valid answer and that less, but more accurate data is better than more, less accurate data.

Using Language Qualifiers

For this project, IEI tried an Italian language test to identify fluent Italian researchers, in the hope that Italian speakers would be able to retrieve the required data accurately, but the response rate was so low as to negate the qualification.

Amazon Mechanical Turk does not currently collect language proficiency information from turkers. The only possibility for getting workers fluent in a particular language is to note the country where a worker is. This doesn't guarantee a turker's language proficiency, though; a native Italian speaker, for example, could very well be found in the U.S.A, India, or any other country with a different local language.

Although it did not pan out this time, the Italian language qualifier test was an interesting experiment and IEI has an excellent working language test template for future use.

About Information Evolution, Inc.

IEI provides human resource and technology services to companies, primarily in publishing or related industries, that manage large databases in real time. For more information, call (512) 650-1111 or visit www.informationevolution.com.