Designing a Scalable Crowdsourcing Platform

Chris Van Pelt
Founder and CTO
CrowdFlower
2111 Mission St. #302
San Francisco, Ca
vanpelt@crowdflower.com

Alex Sorokin
Principle Data Scientist
CrowdFlower
2111 Mission St. #302
San Francisco, Ca
alex@crowdflower.com

ABSTRACT

Computers are extremely efficient at crawling, storing and processing huge volumes of structured data. They are great at exploiting link structures to generate valuable knowledge. Yet there are plenty of data processing tasks that are difficult today. Labeling sentiment, moderating images, and mining structured content from the web are still too hard for computers. Automated techniques can get us a long way in some of those, but human intelligence is required when an accurate decision is ultimately important. In many cases that decision is easy for people and can be made quickly - in a few seconds to few minutes.

By creating millions of simple online tasks we create a distributed computing machine. By shipping the tasks to millions of contributors around the globe, we make this human computer available 24/7 to make important decisions about your data. In this talk, I will describe our approach to designing CrowdFlower - a scalable crowdsourcing platform - as it evolved over the last 4 years.

We think about crowdsourcing in terms of **Q**uality, **C**ost and **S**peed. They are the ultimate design objectives of a human computer. Unfortunately, we can't have all 3. A general price-constrained task requiring 99.9% accuracy and 10 minute turnaround is not possible today. I will discuss design decisions behind CrowdFlower that allow us to pursue any two of these objectives.

I will briefly present examples of common crowdsourced tasks and tools built into the platform to make the design of complex tasks easy, tools such as CrowdFlower Markup Language(CML).

Quality control is the single most important challenge in Crowdsourcing. To enable an unidentified crowd of people to produce meaningful work, we must be certain that we can filter out bad contributors and produce high quality output. Initially we only used consensus. As the diversity and size of our crowd grew, so did the number of people attempting fraud. CrowdFlower developed "Gold standard" to block attempts of fraud. The use of gold allowed us to train contributors for the details of specific domains. By defining expected responses for a subset of the work and providing explanations of why a given response was expected, we are able distribute tasks to an ever-expanding anonymous workforce without sacrificing quality. As the volumes and demands for gold standard data grew, we developed automated techniques to generate gold in unlimited quantities to better train workers and minimize internal human resources that are required to run these jobs. As humans naturally make mistakes, we collect and aggregate multiple judgments to reach our target quality. When tasks are too subjective to specify a "Gold standard," CrowdFlower can fall back onto peer review. Finally, we track historical contributor's performance in a specific domain to reduce the amount of training and evaluation a contributor has to go through in a single task.

When cost is the driving factor in the decision to crowdsource, CrowdFlower provides access to millions of contributors. As we automatically determine who can contribute, we allow absolutely anyone to try the task to see if they can do it. Crowdsourcing differs from traditional service oriented businesses in that costs increase with volume. The list of tasks that a crowd can choose from is an open marketplace. Contributors go towards the best paying or most enjoyable tasks available. I will discuss techniques we use to keep the marketplace as stable as possible.

The speed at which work can be completed is often the primary requirement of our clients. To maximize the scalability of our workforce, we use a channel based approach for our labor partnerships. As a result, we rarely have supply side constraints. This strategy also gives us a worldwide presence to achieve 24/7 processing. Some of our channel partners reward their users with virtual currency or other valuable items. I will briefly address the relationship between game dynamics and crowdsourced work.

As CrowdFlower's workflows continue to become more advanced, we envision a world where humans and machines work together to produce high quality data faster and cheaper; a world, where work is available to everyone on the globe; where job-specific education is provided on demand as it is needed.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Crowdsourcing, Data, Workforce

SIGMOD '12, May 20–24, 2012, Scottsdale, Arizona, USA.
ACM 978-1-4503-1247-9/12/05.