Internet-Scale Collection of Human-Reviewed Data

Qi Su, Dmitry Pavlov, Jyh-Herng Chow, and Wendell C. Baker
Yahoo! Inc
Sunnyvale, CA, USA

ABSTRACT

Enterprise and web data processing and content aggregation systems often require extensive use of human-reviewed data (*e.g.* for training and monitoring machine learning-based applications). Today these needs are often met by in-house efforts or out-sourced offshore contracting. Emerging applications attempt to provide automated collection of human-reviewed data at Internet-scale. We conduct extensive experiments to study the effectiveness of one such application. We also study the feasibility of using Yahoo! Answers, a general question-answering forum, for human-reviewed data collection.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; J.4 [**Social and Behavioral Sciences**]

General Terms

Experimentation, Measurement, Human Factors

Keywords

human data, manual review, data collection

1. INTRODUCTION

Today's enterprise and web content management systems automate the integration of data from heterogeneous sources. For example, Yahoo! Marketplace verticals (*e.g.* Yahoo! Travel, Local and Shopping) aggregate structured as well as unstructured content from paid feed providers, user submission, as well as web crawling. Human-provided data plays a crucial role in the effective operation of such automated systems. Content aggregation usually includes cleansing and enrichment applications such as attribute extraction, categorization, and entity resolution [4], which we will refer to as ACE:

- Attribute extraction is the recovery and labeling of a subcomponent adjectival description such as *128 MB* (capacity), *Size 10* (shoe size), *10:00-6:00 PM* (time), *Pizza in a Cup!* (brand name), etc., from a data record.
- Categorization is the assignment of a data record into one or more locations within an external taxonomy.
- Entity resolution is the identification of relationships between data records that refer to real world entities, *e.g. Hilton San Jose, 123 Main Street* and *SJ Hilton, 123 Main St.* refer to the same hotel.

ACE applications, regardless of whether they are machine learning-based or rule-based, all require human-labeled gold-standard data for bootstrapping (machine learning algorithm training, rule inferencing, *etc.*), quality-assurance monitoring, and feedback or refinement. Furthermore, human review is often critical for controlling data quality: for example, external data feeds may need to be manually reconciled for errors or missing information; user generated content may need to be moderated for spam and offensive content.

There has been research on effectively leveraging human review resources, in particular in the area of active learning, which aims to maximize improvement in the performance of a machine learner using minimal human review resources. However, there has been little study of mechanisms to collect human-reviewed data at a large scale and of the behavior of the respondents providing the data.

In industry, a common human data collection mechanism is the use of low-cost contract workers whose activities are coordinated by a locally-run service bureau. These providers generate human-labeled data for many of the aforementioned ACE applications. This review and labeling ecosystem offers a single convenient business interface to the labeling task, but has scalability limits in the incremental and ongoing relationship cost, as well as the latency and throughput of the integrated supply chain management system that governs it.

As user-generated content proliferates on the web, websites are devoting increasing resources [2] to moderation and abuse detection. These highly-trained, typically in-house, editorial efforts often have high per-employee throughput. However, scalability is limited by high overhead costs such as recruiting, staff turnover and training.

In contrast, systems such as Google Image Labeler [9] and Amazon Mechanical Turk [1] attempt to scale to the Internet audience at large for the purpose of collecting human-reviewed data. Google Image Labeler, based on the ESP Game [20] [23], offers points with no monetary value as a reward for image tagging tasks with a head-to-head competition model to encourage good answers. Mechanical Turk provides monetary rewards (at least one cent) for tasks, can be used for any generic task, and there is no explicit collaboration between answerers. Beyond these two systems, we

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

can also view general question-answer venues, such as Yahoo! Answers [24], Usenet or discussion forums, as potential technology platforms for the collection of human-reviewed data.

Collecting human-reviewed data at Internet-scale has the potential for breaking the throughput bottleneck of in-house or outsourced providers while lowering the cost-per-unit of high-quality human review. For example, it is suggested [20] that an ESP game system with 5000 active users at any given time would be able to label Google's image index in weeks. Two months after launch, Google Image Labeler [9] shows that the top five users have individually labeled over 8,000 images. Google does not incur any explicit per-unit cost in collecting this data, other than the overhead of operating the site. The ESP study [20] shows good results on several quality metrics. In another application, as part of the search for Jim Gray [18], 560,000 satellite images were reviewed by volunteers in 5 days.

In this paper we conduct what we believe is the first public study of an Internet-scale human-reviewed data collection mechanism focusing on data quality, task throughput, and user behavior. We also conduct experiments on Yahoo! Answers to study the feasibility of using general question-answering forums for human-reviewed data collection. We survey related work in Section 2. Section 3 presents an overview of the system we are studying as well the tasks and datasets used in our experiments. Section 4 discusses the design and results of our experiments. We repeat some of our experiments on Yahoo! Answers in Section 5. We conclude in Section 6 with substantial evidence that high-quality human-reviewed data can be acquired at low cost at Internet-scale.

2. RELATED WORK

2.1 Active Learning and Collaborative Filtering

Supervised machine learning algorithms are commonly used in many ACE applications. Having good labeled training datasets is crucial to the performance of these learning algorithms, but using human review to label data is generally expensive. Active learning [6] focuses on maximizing the return on limited human review resources. Algorithms attempt to pick the subset of the unlabeled data such that when labeled it will provide maximal improvement to the learner. Active learning has been applied to a number of application areas, including entity resolution [16], information extraction [3], and text classification [13]. Collaborative filtering [5] [15] focuses on using a collection of user-provided or generated data to make predictions about user preferences. A common application is recommendation systems for user tastes in subjects such as books, music, movies, etc. Our work is complementary to both research areas, as we focus on the mechanisms that provide manual review resources.

2.2 Data Cleaning

Data cleaning (see [7] [10] [14] for a sample of papers), also known as data quality, data cleansing, is a research area which studies the detection and removal of data errors and inconsistencies in order to enhance data quality. Entity resolution [4], which is one of the applications in our experiment, is a common data cleaning application that identifies relationships between data records that refer to real world entities. Data cleaning research has largely focused on algorithmic solutions and data quality metrics. Our work focuses on mechanisms to efficiently collect human-reviewed data than can be consumed by these algorithmic solutions to enhance data quality.

2.3 Human Computation

Human Computation [19], related to Crowdsourcing [11], is an emerging paradigm of leveraging masses of humans to perform computation or to create content [12]. Surowiecki [17] characterizes behavior of human collectives and presents effective collective decision-making applications. Three games developed at Carnegie Mellon University, ESP Game [20], Peekaboom [21], and Verbosity [22], are exemplary applications of human computing. The ESP Game asks two random users to collaboratively tag an image, completing successfully when both users provide a common tag. Peekaboom asks one user to reveal a portion of an image to induce the collaborating user to guess a given word. User data is used to isolate the portion of the image corresponding to a word. Verbosity collects common sense human knowledge facts from users. The system that we study is a more general application of human computation. Our work focuses on the analysis of the collected human-reviewed data rather than the design of novel interfaces used to collect data. Gentry *et al.* [8] analyze human computation from a security and reliability perspective.

3. SYSTEM M TASKS

In this study, our primary experiment platform is a web-based human data collection system that we will refer to by the alias System M. Unlike Google Image Labeler, System M is open to the public to submit tasks to be answered. Our result is a novel use of System M. Our experiments replicated the current data supply chain in the form of tasks on System M. System M has a mature implementation, provides structured response interface, and offers many useful features for our human-reviewed data collection applications. However, it has a limited user base and worker rewards are monetary (rather than reputation or points, which would decrease requester cost). In contrast, general question & answer forums such as Yahoo! Answers have much larger user base and worker rewards are non-monetary reputation or points, though the collected data tends to be unstructured. In Section 5, we repeated some of the System M experiments on Yahoo! Answers.

3.1 System M Overview

On System M, requesters post tasks for registered workers to answer. For a given task, the requester specifies the lifetime that the task is available to be accepted by workers, the length of time a worker has to complete the task once he accepts it, the number of answers to accept (guaranteed to come from distinct workers), the reward per approved answer, and optionally, the qualifications of the workers who are eligible to answer the task. Tasks can be extended at any time by extending the lifetime and/or increasing the number of answers to accept. The requester processes the answers and unilaterally decides to approve or reject each answer. Note only approved answers are paid to the worker. Currently the system is US-based, as requesters and workers are required to link a US-based bank account. The system

provides REST and SOAP APIs to requesters for automated interactions with the system.

System M allows requesters to specify restrictions on qualifications that a worker must have in order to be eligible to answer a given task. The built-in qualification types include system-tracked statistics of the worker such as percentage of tasks accepted, completed, approved, rejected, *etc*, as well as worker locale. Requesters can also create custom qualification types. The requester who creates a given custom qualification type is responsible for granting initial qualification scores to workers who request this qualification. The requester can update qualification score for any worker at any time. A custom qualification type may optionally include a test, which is an unpaid task. Currently, most requesters who use custom qualifications grant a default score to any worker who requests it. The requester posts tasks that require a minimum qualification score that is lower than the default, thereby allowing any new worker to work on the requester's tasks. Then as a new worker submits answers over time, the requester adjusts the worker's qualification score according to the worker's performance. In our experiments, we grant qualification score as the worker's accuracy score on the corresponding qualification test of 20-21 questions and do not adjust the score once it is set.

Operationally one can observe that a wide range of tasks have been submitted to System M by third party requesters. For example, for two cents, workers are asked to look at the scanned image of a mortgage document to extract 13 data fields. For 68 cents, workers are asked to transcribe an eight minute podcast. For a cent, workers are asked to judge the relative relevancy of two results to a product search. The aforementioned mortgage extraction task is tedious and the reward is very low. However, we still observe the number of available tasks of that type decreasing steadily over time, indicating workers are consuming the tasks. From the requester perspective, this is a encouraging sign that System M users are willing to do a significant amount of work for relatively low rewards. In Section 4.2.3 we will analyze worker pay rates in our experiments.

3.2 Tasks and Datasets

Our experiments focus on two ACE applications: attribute extraction and entity resolution. For entity resolution, we selected resolution of "Yellow Page"-style business data records for hotels. Given two data records with a business name, location and phone information, workers are asked to select the relationship between the two (same business, different business, same location but different name, *etc*). For attribute extraction, we selected the extraction of an age category (adult or kids), brand, or model from a product description text string. Among these four applications, hotel resolution and age extraction problems are categorical valued: workers are asked to choose from a limited set of answers; product brand and model extraction problems require free-text responses. We chose these four applications as they are representative of the content aggregation applications within the Yahoo! Marketplace backend data processing systems where extensive human-labeled data is required.

The datasets used for experiments of these four applications are uniform random samples of human-labeled gold-standard datasets that come from both in-house manual review as well as the external labeling ecosystem.

Task Type	Workers	Qual Test Accuracy Mean	Qual Test Accuracy Stdev
ER Hotels	397	0.91	0.081
Extract Age	275	0.428	0.105
Extract Brand	170	0.706	0.079
Extract Model	80	0.345	0.109

Table 1: Qualification Test Results

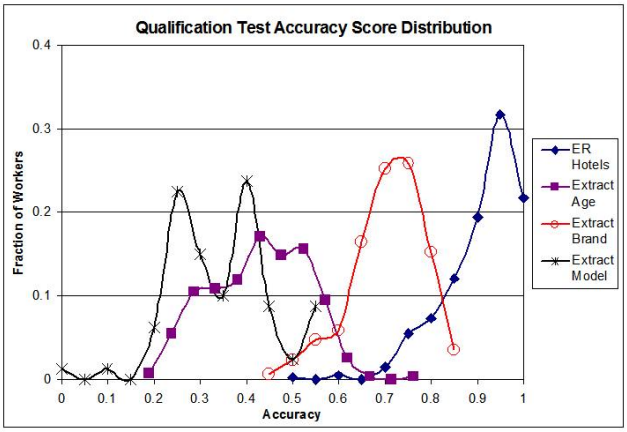

Figure 1: Qualification Scores Distribution

4. EXPERIMENTS

4.1 System M Qualification Tests

As an initial experiment, we posted a qualification test on System M for each of the four applications: hotel resolution, age extraction, brand extraction and model extraction. The qualification tests are taken voluntarily by workers and are unpaid. We included twenty questions in each qualification test, except the age test, which had 21 questions.

In the hotel resolution test, for each of 20 pairs of hotel business records (with name, location, and phone information) given, the worker was asked to choose among five choices: same business, different names but same location, same business but different location, two completely unrelated businesses, or other. For age, brand or model extraction, the worker was given 20 (brand and model) or 21 (age) product description text strings and asked about each. For age extraction, the worker was asked to select from three categories: adult, kids or not applicable. For the brand and model extraction tests, the worker was asked to type in the product brand or model, respectively. Note the question format for the general experiments in Section 4.2 was the same as that of the qualification test, except that the qualification test had 20 or 21 questions, whereas each general experiment task is comprised of a single question.

Table 1 shows the number of participating workers and accuracy results on the four qualification tests. There was significantly more participation in the multiple-choice task types (hotel resolution and age extraction) than the free text task types. Figure 1 delineates the distribution of accuracy scores among the workers who took each qualification test. Table 1 shows that, on average, workers performed best on hotel resolution, followed by brand extraction. Accuracy is

Task Type	Qual Accuracy Required	Answers per Task	Voting Threshold	Eligible Workers	Particip. Workers	Elapsed Time	Answer Accuracy	Voted Answer Accuracy	Voted Answer Adjusted Accuracy
ER Hotels	1.0	3	2	69	14	361min	0.764	0.82	0.851
ER Hotels	0.9	3	2	245	20	68min	0.774	0.817	0.842
ER Hotels	None	3	2	All	32	53min	0.76	0.777	0.798
ER Hotels	None	5	3	All	42	111min	0.735	0.75	0.781
Age	0.57	3	2	27	10	1099min	0.863	0.947	0.95
Age	0.43	3	2	165	18	142min	0.726	0.77	0.802
Age	None	3	2	All	22	42min	0.944	0.977	0.977
Brand	0.75	3	2	72	12	124min	0.768	0.847	0.904
Brand	None	3	2	All	29	37min	0.676	0.69	0.796
Model	0.35	3	2	32	8	146min	0.727	0.8	0.851
Model	None	3	2	All	23	43min	0.681	0.703	0.851

Table 2: Task Results

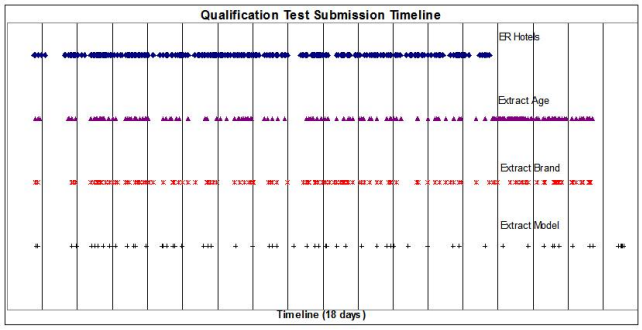

Figure 2: Qualification Test Submission Timeline

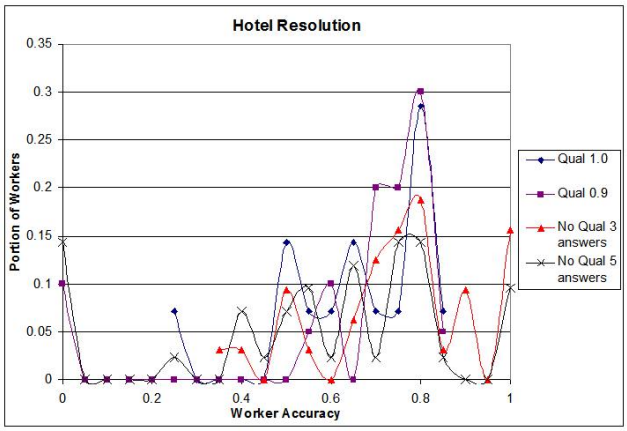

Figure 3: Worker Accuracy Distribution - Hotel Resolution

significantly lower on age and model extraction qualification tests. Figure 2 shows a horizontal timeline for submission of qualification tests by workers. The starting point of the graph is midnight October 31, 2006, Eastern Time. Each vertical gridline is midnight Eastern Time. The four qualification tests were posted on System M at 5:13PM Eastern Time on October 31. Note the qualification tests were deactivated on System M after the general experiments were done. The time series for the hotel resolution qualification test ends the earliest because the hotel resolution experiments were completed first.

4.2 System M Experiments

4.2.1 Experiment Settings

After we obtained preliminary distribution of accuracy scores from the qualification tests, we selected qualification score cutoffs to test the efficacy of having qualified versus unqualified workers answering each application. We conducted 11 experiments (we will also refer to them as task types), one per row on Table 2, thus there were four experiments for hotel resolution, three for age extraction, and two each for brand and model extraction. Each experiment was comprised of 300 distinct tasks. Each task was a single question: asking the worker, for example, to select the relationship between a pair of hotel business data records, or to extract product brand from an unstructured product description text string. These tasks were paid one cent each.

We stated in the question overview text that the worker's answer will be approved (hence we pay the reward to the worker) only if the majority of workers who answered the question agree. Taking into account the half cent commission to System M, our net cost per task is up to 4.5 cents (when we collect 3 answers) or 7.5 cents (5 answers). Each experiment had the following parameters: one of the four applications; the qualification required for worker to be eligible to work on the task: the minimum accuracy required on the corresponding qualification test, or none for the case where every worker is eligible; the number of answers to collect per task, with each answer guaranteed by System M to come from distinct workers; the voting threshold (the number of agreeing answers required to declare a voted answer as valid). For each task, System M collects the number of answers specified. We examine the answers to see if there is a voted answer meeting the threshold number of votes. If so, those with the voted answer are approved and paid, while the rest of the answers are rejected without pay. If there is no voted answer, then all answers are rejected without pay. The approve/reject decision is made and the workers are notified in batch after all 300 tasks of each task type have been answered.

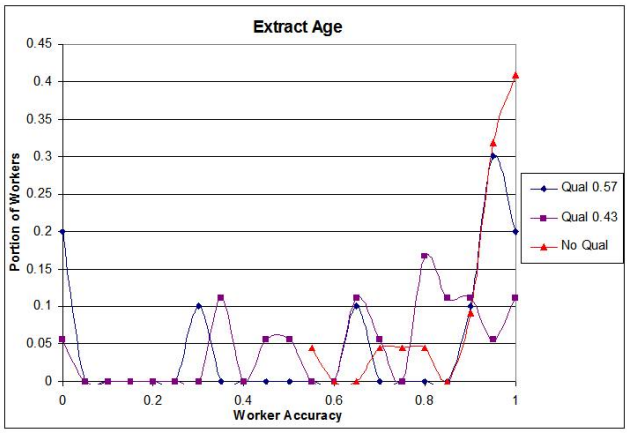

Figure 4: Worker Accuracy Distribution - Age Extraction

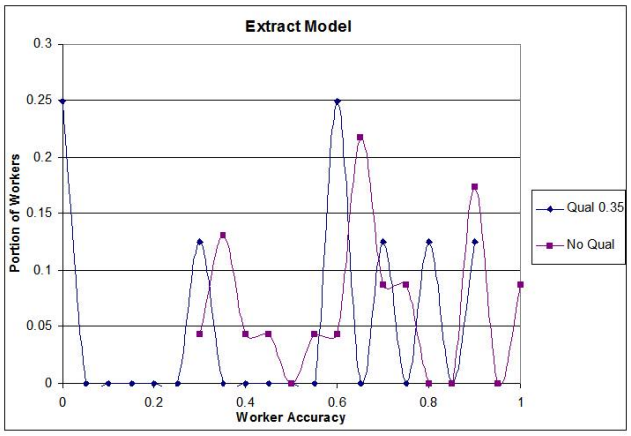

Figure 6: Worker Accuracy Distribution - Model Extraction

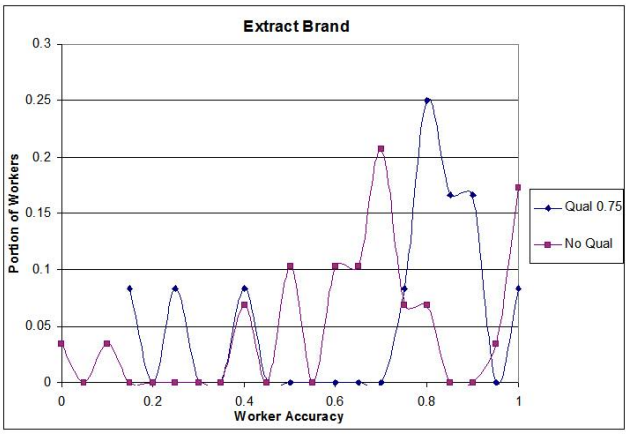

Figure 5: Worker Accuracy Distribution - Brand Extraction

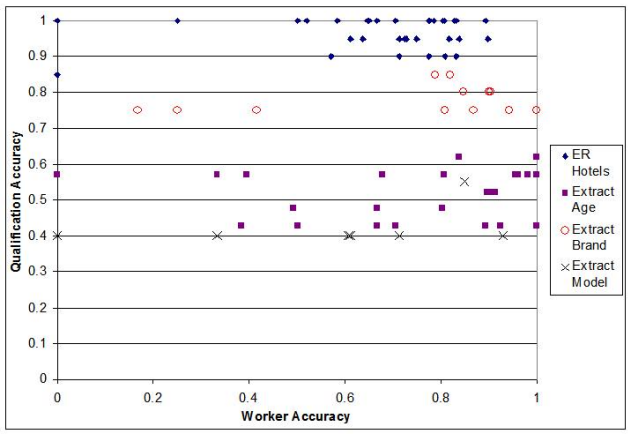

Figure 7: Worker Accuracy vs. Qualification Accuracy

For each of the four applications, we conducted at least one experiment with a qualification restriction and at least one experiment without any qualification restriction in order to determine whether there is an accuracy difference between the presence or absence of qualification restrictions, and between higher and lower qualification cutoffs. For the hotel resolution problem, we conducted a separate experiment that collected five answers per task to contrast with the collection of three answers per task in order to see if collecting more answers per task leads to higher accuracy of the voted answer.

4.2.2 Accuracy Results

Table 2 shows an overview of the results of the experiments. For each task type with qualification requirements, the number of eligible workers (during the time the tasks were available) is recorded. For each experiment, the table lists the number of distinct participating workers, the elapsed time until every question has collected the requisite number of answers (3 or 5). The "Answer Accuracy" column indicates the fraction of all worker answers (900 or 1500) that were correct, according to our gold standard labeled dataset. The "Voted Answer Accuracy" column shows the fraction of the 300 tasks where the voted answer (agreed by at least 2 out of 3 or at least 3 out of 5 answers) was correct. The "Voted Answer Adjusted Accuracy" figure is computed by taking the number of tasks with correct voted answer divided by the number of tasks with either correct or incorrect answers, thereby ignoring those tasks without a voted answer because no answer had the necessary number of votes.

Figures 3, 4, 5, and 6 show discretized distributions of worker accuracy (aggregate accuracy of all the answers submitted by a worker on a given task type experiment) with one graph per application. The worker accuracy scores are discretized into intervals of width 0.05 ([0,0.05), [0.05,0.1).., [0.95,1),[1,1]). The y-value (fraction of workers with accuracy in a given interval) is plotted against the lower bound of the interval. Figure 7 plots worker accuracy versus the worker's accuracy score on the corresponding qualification test. Multiple experiments for the same application are combined into a single data series.

Table 2 shows that, for each application, as we lower or remove the qualification requirement, the number of participants increases and the elapsed time to complete all the tasks decreases. In all experiments, the majority threshold

	All Participants			Participants Answering >=5% of Tasks			2 Highest Paid Participants Answering >=5% of Tasks		
Task Type	**Workers**	**Hourly Pay**	**Answer Accuracy**	**Workers**	**Hourly Pay**	**Answer Accuracy**	**% Tasks Answered**	**Hourly Pay**	**Answer Accuracy**
ER Hotels qual 1.0, 2/3	14	$0.89	0.764	10	$0.89	0.772	54% 48%	$2.45 $2.43	0.828 0.804
ER Hotels qual 0.9, 2/3	20	$0.91	0.774	12	$1.20	0.78	89% 12%	$2.93 $1.71	0.785 0.838
ER Hotels no qual, 2/3	32	$1.10	0.76	13	$1.30	0.759	23% 14%	$2.31 $2.20	0.786 0.829
ER Hotels no qual, 3/5	42	$0.97	0.735	19	$1.10	0.746	5% 9%	$2.67 $2.42	0.438 0.778
Age qual 0.57	10	$0.78	0.863	5	$0.78	0.872	93% 100%	$4.31 $4.06	0.828 0.98
Age qual 0.43	18	$1.77	0.726	10	$1.86	0.724	41% 17%	$4.90 $3.45	0.839 0.9
Age no qual	22	$3.06	0.944	10	$3.16	0.946	79% 63%	$6.53 $5.52	0.97 0.963
Brand qual 0.75	12	$1.39	0.768	10	$1.39	0.77	12% 30%	$2.47 $2.24	0.943 0.82
Brand no qual	29	$1.44	0.676	13	$2.06	0.676	74% 11%	$5.82 $4.94	0.726 0.719
Model qual 0.35	8	$1.35	0.727	5	$1.68	0.734	59% 43%	$4.02 $2.01	0.848 0.612
Model no qual	23	$1.62	0.681	18	$1.62	0.68	8% 16%	$3.49 $2.70	0.917 0.653

Table 3: Worker Productivity Results

voting scheme (2/3 or 3/5) resulted in higher accuracy than the average accuracy of the underlying answers, thereby demonstrating wisdom of the crowd. The two hotel resolution experiments without qualification but with different voting schemes showed minimal difference in accuracy, suggesting that accuracy is boosted as long as there is some kind of voting.

In terms of correlation between answer accuracy and qualification requirement, we had expected that requiring a higher qualification accuracy would lead to higher answer accuracy than requiring a lower qualification accuracy, and that having a qualification requirement would lead to higher answer accuracy than not having one. The hotel resolution, brand and model extraction experiments validated this hypothesis. However, the age extraction experiment without qualification requirement was the only outlier. It showed much higher answer accuracy than the two age experiments with qualification requirements. We believe this is the result of having prolific participating workers who are by chance significantly above-average in accuracy. Figure 4 shows that most of the distribution of workers for the age extraction without qualification experiment is concentrated at very high accuracy region between 0.9 and 1. Turns out these high accuracy workers also were very prolific, therefore biasing the average answer accuracy upward. In fact, the most prolific 20% (4 out of 22) of workers provided 69% of all the answers at 97.1% accuracy. With only 10s or 20s of participating workers in most of the experiments, a few prolific workers with above or below average accuracy could bias the average accuracy on an experiment. We believe our hypothesis does hold, though larger scale experiments are needed. Overall, the accuracy achieved with qualification requirements on the age, brand and model extraction was very encouraging for practical applications. The accuracy on the hotel resolution experiments was lower than expected.

The worker accuracy distributions shown in Figures 3, 4, 5, and 6 look very different from the worker qualification accuracy distributions in Figure 1. Brand is the only similarity, where both qualification and experiment accuracy distributions have most of the mass around accuracy of 0.7 to 0.8. Compared to the corresponding qualification accuracy distribution, the hotel resolution experiment accuracy distributions are shifted toward lower accuracy, while age and model experiment accuracy distributions are shifted toward higher accuracy. Figure 7 indicates that in general, there seem to be little correlation between qualification and experiment accuracy. For hotel resolution, age and model extraction, for similar qualification accuracy, the worker accuracy in the experiments varies from 0 to 0.8. The brand extraction data series suggests a correlation, with its main cluster around worker accuracy range of 0.8 to 1, though there are three outliers with low worker accuracy.

4.2.3 Worker Productivity Results

Figure 8 shows, for each of the 11 experiments, the fraction of all collected answers that are provided by the top 10% of participating workers. It suggests that the distribution of workers and the number of responses that they submit is not uniform. In all cases, a minority of prolific workers submitted a disproportionate amount of answers. For most experiments, the most prolific 10% of workers submitted 20% to 40% of the answers. In one experiment, the top 10% submitted over 80% of the answers. We can look at Figure 9 ignoring the x-axis to see that the long-tail distribu-

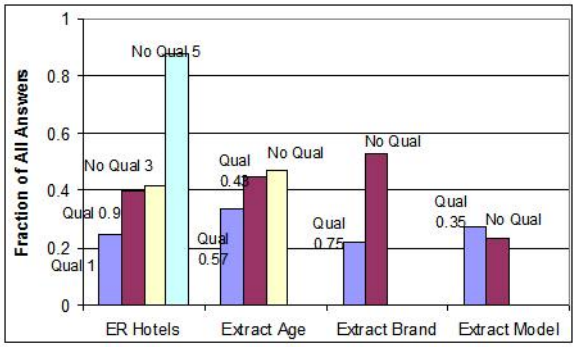

Figure 8: Fraction of All Answers Provided by Top 10% of Workers

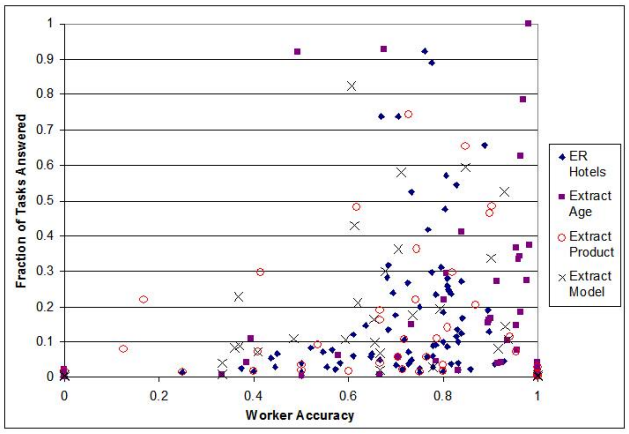

Figure 9: Worker Accuracy vs. Fraction of Tasks Answered

tion is evident for all four applications. A few workers were very prolific, while a majority of the workers only answered less than 10% of the questions, as shown by the concentration of data points having y-value less than 0.1. Figure 9 plots worker accuracy on the x-axis against a measure (the fraction of the 300 tasks answered by the worker) of how prolific the worker is on the y-axis. Multiple experiments for the same application are combined into a single data series. The general trend is that the less prolific workers had a much great variance in accuracy than the more prolific workers. For workers who answered more than 20% of the questions, on the age extraction experiments, most had over 0.9 accuracy, with two outliers answered over 90% of the questions but had about 0.5 and 0.7 accuracy. For the other three applications, the workers who answered more than 20% of the questions generally were in the accuracy range of 0.6 to 0.9.

We analyzed the timing of worker submissions to compute hourly pay rates for the workers. Each answer submission comes with an accept timestamp (when the worker accepted the task to work on it) and an submit timestamp (when the worker completed the task and submitted the answer). Given all the answer submissions from a particular worker on a given task type, we derive the time spent by the worker as the latest submit timestamp subtracted by the earliest accept timestamp. The hourly pay rate is computed as the

	#Task Types						
	1	**2**	**3**	**4**	**5**	**6**	**Total Workers**
All	110	27	10	5	2	1	155
Answering >=5% Tasks	67	16	3	3	1	0	90

Table 4: Worker Participation in Task Types

number of answers approved by the threshold voting scheme multiplied by one cent (reward per task) divided by the time spent. Note this computation assumes that the worker is not working on other task types (e.g. from other requesters) concurrently.

Table 3 shows, for each task type, the average worker hourly pay rates and the two highest individual hourly pay rates along with the corresponding worker answer accuracy. For each task type, the average hourly pay rate for a group of workers is computed by dividing the aggregate reward paid divided by the aggregate time spent. For each task type, we compare the average hourly pay rate and answer accuracy for all workers, prolific workers (defined as those who answered at least 5% of the 300 tasks), and the two prolific workers with the highest hourly pay rate. For each of the four applications (hotel resolution, age/brand/model extraction), the average pay rate of all workers increase as the qualification becomes less restrictive. As we had discussed in Section 4.2.2 and we can see in the Answer Accuracy column for All Participants in Table 3, overall accuracy tends to decline as the qualification becomes less restrictive. A possible explanation of the pay trend is that as the qualification becomes less restrictive, we see more participants who are not as serious about work quality and are answering the tasks quickly to make money. Comparing pay rate of the prolific workers against all workers, of the 11 task types, in seven cases the prolific workers are paid more and in the rest, the pay is the same. Comparing answer accuracy, in seven cases the prolific workers are more accurate, and in the rest, the prolific workers are just as accurate or marginally less accurate. The rightmost three columns of Table 3 show that the highest paid workers earn significantly more than the average pay rate. They also tend to be prolific (average 41% of tasks answered) and accurate (only 4/22 have lower than average answer accuracy). The top paid prolific worker of all the experiments was someone on the age extraction without qualification task type who averaged $6.53 per hour. This worker answered 236 age extraction tasks within 21 minutes, averaging 5.4 seconds per task, at an above average 97% accuracy. Overall, we see that the vast majority of the workers participating in the experiments earned significantly below minimum wage rates, while very few prolific workers on a subset of the task types approached the minimum wage rate.

4.2.4 Worker Accuracy Across Task Types

Table 4 shows the number of workers who participated in one or more task types. The last row qualifies the results to count participation as a worker answering at least 5% of the tasks of a given task type. In total, of the 155 distinct workers who participated in our 11 task types, 45 of the workers participated in more than one task types. If we define participation as a worker answering at least 5% of

Task Type	ER Hotels	Extract Age	Extract Brand	Extract Model
Avg Answers/ Question	2.2	3.3	4.1	1.8
Questions Answered	100%	90%	95%	95%
Questions w/ Useful Answers	90%	86%	90%	65%
Questions w/ Parsable Answers	80%	86%	90%	30%
Questions w/ maj. Useful Answer Correct	45%	24%	50%	20%
Questions w/ maj. Parsable Answer Correct	45%	29%	55%	15%

Table 5: Yahoo! Answers Results

the tasks for a task type, 90 distinct workers participated in the 11 task types, 23 of whom particpated in more than one task type.

We drill down further to examine answer accuracy of each worker across task types. We do not see a clear correlation in worker performance across task types. Some workers are consistently accurate across task types. For example, a worker had 0.94 answer accuracy answering 12% of tasks on brand extraction with 0.75 qualification, 0.98 accuracy on 100% of tasks on age extraction with 0.57 qualification, and 0.96 accuracy on 33% of tasks on age extraction with 0.43 qualification. This worker was very prolific with significantly above-average answer accuracy across all three task types. Of course we also have workers who consistently showed answer accuracy below the task type average, as well as workers with mixed performances on task types. For example, in the preceding section, we mentioned the worker who averaged $6.53 per hour with above-average 97% accuracy on the no qualification age extraction task type. However, this worker showed below average accuracy on three other task types: 61% accuracy answering 43% of tasks on the model extraction with 0.35 qualification task type; 73% accuracy answering 27% of tasks on the hotel resolution with 0.9 qualification; and significantly below average 49% accuracy answering 92% of tasks on age extraction with 0.43 qualification. Of the 45 workers who participated in multiple task types, 13 (29%) showed above-average answer accuracy on all task types, while 7 (16%) showed below-average accuracy on all task types. Considering participation as a worker answering at least 5% of tasks, of the 23 workers who participated in multiple task types, 9 (39%) had above-average answer accuracy on all task types, while 5 (22%) had below-average accuracy on all task types. We manually reviewed the answers provided by the aforementioned 5 workers with consistently below-average answer accuracy and did not find any patterns of dishonest or malicious behavior.

5. YAHOO! ANSWERS

Having seen encouraging results from the user community of System M, can we leverage large existing online social network sites for the collection of human-reviewed data? The first candidate that comes to mind is Yahoo! Answers. Yahoo! Answers is a general question-answer discussion forum organized by a topic hierarchy. In the 11 month since launch, it has accumulated 65 million answers [24]. It is similar to Google Image Labeler in that there is a base of dedicated users (two top users have each answered over 50,000 questions) and the system does not provide monetary reward to users. Two questions come to mind: can we engage the Answers user community to participate in our application? Can the underlying Answers technology platform be leveraged for our application? To answer these questions, over a period of four days, we manually submitted the 81 questions from the four qualification tests from Section 4.1 to Yahoo! Answers. Each question was available to be answered by users for three days.

We received encouraging number of answers: 95% of questions had at least one answer. We manually reviewed each answer to label whether it is spam, whether it is useful and answers the question, whether it is machine parsable, and whether it is correct. Table 5 shows the results of the experiments in detail. The last two rows show the accuracy if we use majority voting on the useful (resp. parsable) answers of each question. A few data points are not in the table: about 1% of the answers were spam, 4% were unhelpful responses such as "who cares?" and about 13% of the answers did not directly answer the question (*e.g.* one answered "It's an Air Conditioner" for a brand extraction question). Of interest here are the 8% of answers which contained extra useful information beyond what was asked in the question. For example, on one hotel resolution question, a user actually called the phone number in the question data to determine that the two hotel records presented were the result of a franchise change at the location. On hotel resolution questions, 45% of answers were correct. On extraction questions, age had a 28% answer accuracy, brand had a 51% answer accuracy and model had a 17% answer accuracy.

The biggest challenge with using a discussion forum like Yahoo! Answers for automated collection of human-reviewed data is parsing. The challenge is two fold: separating the useful data from the spam and unhelpful data; secondarily the parsing out of the user's intended response from pleasantries and grammatical "glue." The challenge is largely a consequence of user behavior. Forums like Yahoo! Answers are meant for human exchanges, hence users are used to receiving conversational questions and responding with breezy and off-the-cuff answers. One imagines that this sort of natural language give-and-take provides users with a dimension of confidence in the interrogator and respectively the responder that can only be assessed by a living person. For example, our hotel resolution questions look like the following:

```
...
What do you think is the relationship between the
two businesses described by the two records? Is it:
A. The two records are about the same business.
B. The two records have different names but are at
the same location.
...
```

Of course, it is trivial to parse the verbatim answers, which were the result of cut-and-paste in the browser. The difficulty comes, often, when the user answers were fluent conversational responses, such as: "Since you didn't give your sources, I am inclined to answer with 'F' " or "My guess is C because they are both hotels." On the free text

questions such as brand or model extraction, simple regular expression templates could potentially handle terse responses such as "The brand is Panasonic" or "It's made by Pickett." One can't expect to exhaustively list all such possible text patterns, so this approach has clear limitations. In some cases, we needed to segment the answer into sentences to filter out the irrelevant statements; for example, "It's a BOSCH. We have a BOSCH and it works great!" In some cases the user provided multiple answers, for example, "I'd say either Sanford or Prismacolor??" . The double question marks indicate responder uncertainty which complicates the answer recovery problem.

In contrast, on System M, multiple choice questions did not present any ambiguity in the divination of user intent as the user was choosing radio buttons in the browser GUI. For free text questions on brand and model extraction, none of the System M workers entered extraneous text. Clearly, the user behavior is very different on System M, as workers are doing a task for the requester in an explicit paid relationship, rather than having a potentially open-ended *pro bono* question & answer-type conversation with a fellow user. The System M workers expect to be evaluated on their answers and there is explicit monetary payment associated with this evaluation; in contrast, the question & answers system is informal and answerers accrue irredemable, non-monetary "points."

On account of the large user base alone, Yahoo! Answers is a promising vehicle for automated collection of human-reviewed data. We saw decent participation from users. On two of the task types (hotel resolution and brand extraction), we saw reasonable answer accuracy of over 45%. The challenge of user answer parsing can be mitigated in several ways with small changes to the underlying technology infrastructure: support multiple choice questions which can be modeled as polls; as part of the question text, explicitly state the answer collection is automated and ask that users do not type in extraneous text; implement clever user interfaces (*e.g.* for brand/model extraction, require the user to select a substring from the product description text). Applications of general question-answer forums such as Yahoo! Answers, as well as user behavior in such venues, deserve significant research attention.

6. CONCLUSIONS

In this study, we conducted experiments analyzing the data quality, throughput and user behavior of an Internet-scale system for the collection of human-reviewed data. The tasks we experimented with were real content aggregation applications using real-world data. The main contributions of our work are the detailed study using real datasets and the thorough analysis of the resulting data quality and user behavior. Our results show that by applying worker prequalification mechanisms, we are able to obtain an 82% accuracy on hotel resolution, 95% accuracy on product age category extraction, 85% accuracy on product brand extraction and 80% accuracy on product model extraction. These quality measures are very encouraging for a wide variety of practical ACE applications, from creating labeled training sets for machine learning algorithms to providing labeled datasets for quality assurance monitoring. We extend discussion of applications of human-reviewed data in Section 6.1. We envision future enterprise and web information integration and content aggregation systems will include wrappers to interface with Internet-based human-reviewed data collection systems, such that the data processing system can push human review requests to the data collection system on demand.

In terms of future work, we would like to investigate more human-reviewed data collection systems and incentive schemes, as well as conduct larger scale experiments with data from more human-data consuming applications. Are some types of tasks more suitable than others for large scale human review? On unstructured systems such as Yahoo! Answers, we would like to study techniques to parse responses. We would also like to study richer interfaces for general users as a solution to the parsing challenge. For example, can the head-to-head collaborative paradigm of the ESP Game be applied to other types of tasks, such as entity resolution or attribute extraction? Another emerging application of interest is feedback and suggestion systems such as Yahoo! Suggestion Board [25]. Like Yahoo! Answers, it poses challenges such as interface design and algorithmic solutions to automatically filter out noise and parse responses.

6.1 Data Validation Application

As discussed in Section 1, human-labeled data is very important for many ACE applications, since humans are the only authoritative source for label data. However, having been sourced from fallible humans makes the label data itself imperfect; a given human label could be incorrect, relative to the universal truth (as opposed to the "labeled" truth), for a variety of reasons. Complicating the picture is the condition that so called "gold standard" datasets often have but a single data point per label for reasons of efficiency. Furthermore, some labels are inherently ambiguous and subject to interpretation or the relevant context may be missing information for an accurate labeling. For example, in our experiments on System M, given the product description text "Lakai Men's Soca 2 Shoe", two workers answered that the model is "Soca 2", while one worker answered "Soca". Ignoring the gold standard label, in this case it is difficult to determine which is correct given just the product description. The multiple human data inputs merely provide good candidates for valid model names rather than a definitive answer. For the product description text "adidas Piccolo IV Infants & Toddlers", the so called gold standard model label is "adidas Piccolo IV Infants & Toddlers", which is clearly incorrect since "adidas" is a brand name. In contrast, the model label voted by the workers was "Piccolo IV", which seems correct. In this case, the collected external human-labeled data can serve to correct our internal human-labeled gold standard dataset.

Humans can be used in a feedback loop to validate previous generations of human-reviewed data, resulting in enhanced data quality and reliability. For instance, on System M there are survey-style tasks asking workers to list their top 3 travel destinations. The same requester has a separate set of tasks to validate those answers, asking workers "are x, y, and z valid travel destinations?"

7. REFERENCES

[1] Amazon Mechanical Turk. `http://www.mturk.com/`.

[2] J. Angwin. On the offensive - a problem for hot web outfits: Keeping pages free from porn. *Wall Street Journal*, May 2006.

[3] S. Argamon-Engelson and I. Dagan. Committee-based sample selection for probabilistic classiers. *Journal of Artificial Intelligence Research*, 1999.

[4] O. Benjelloun, H. Garcia-Molina, H. Kawai, T. Larson, D. Menestrina, Q. Su, S. Thavisomboon, and J. Widom. Generic Entity Resolution in the SERF Project. *IEEE Data Engineering Bulletin*, June 2006.

[5] J. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proc. of Uncertainty in Artificial Intelligence*, 1998.

[6] D. Cohn, L. Atlas, and R. Ladner. Improving generalization with active learning. *Machine Learning*, 1994.

[7] H. Galhardas, D. Florescu, E. Simon, C. Saita, and D. Shasha. Declarative data cleaning: Language, model, and algorithms. In *Proc. of VLDB*, 2001.

[8] C. Gentry, Z. Ramzan, and S. Stubblebine. Secure distributed human computation. In *Proc. of ACM Conference on Electronic Commerce*, 2005.

[9] Google Image Labeler. `http://images.google.com/imagelabeler/`.

[10] J. Hipp, U. Guntzer, and U. Grimmer. Data quality mining -making a virtue of necessity. In *Proc. of SIGMOD DMKD Workshop*, 2001.

[11] J. Howe. The rise of crowdsourcing. *Wired*, June 2006.

[12] A. Koblin. The sheep market: Two cents worth. Master's thesis, UCLA, 2006.

[13] A. McCallum and K. Nigam. Employing em in pool-based active learning for text classication. In *Proc. of ICML*, 1998.

[14] E. Rahm and H. Do. Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin*, December 2000.

[15] P. Resnick and H. Varian. Recommender systems. *Communications of the ACM*, March 1997.

[16] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proc. of ACM KDD*, 2002.

[17] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, 2004.

[18] Tenacious Search. `http://openphi.net/tenacious/`.

[19] L. von Ahn. Games with a Purpose. *IEEE Computer Magazine*, June 2006.

[20] L. von Ahn and L. Dabbish. Labeling Images with a Computer Game. In *Proc. of ACM CHI*, 2004.

[21] L. von Ahn, S. Ginosar, M. Kedia, R. Liu, and M. Blum. Peekaboom: A Game for Locating Objects in Images. In *Proc. of ACM CHI*, 2006.

[22] L. von Ahn, M. Kedia, and M. Blum. Verbosity: A Game for Collecting Common-Sense Facts. *ACM CHI Notes*, 2006.

[23] L. von Ahn et al. The ESP Game. `http://www.espgame.org/`.

[24] Yahoo! Answers. `http://answers.yahoo.com/`.

[25] Yahoo! Suggestion Board. `http://suggestions.yahoo.com/`.