Crowdsourcing the Assembly of Concept Hierarchies

Kai Eckert
KR & KM Research Group
University of Mannheim,
Germany
kai@informatik.uni-mannheim.de

Mathias Niepert
KR & KM Research Group
University of Mannheim,
Germany
mathias@informatik.uni-mannheim.de

Christof Niemann
University Library
University of Mannheim,
Germany
christof.niemann@bib.uni-mannheim.de

Cameron Buckner
Department of Philosophy
Indiana University, USA
cbuckner@indiana.edu

Colin Allen
Department of History and Philosophy of Science &
Program in Cognitive Science
Indiana University, USA
colallen@indiana.edu

Heiner Stuckenschmidt
KR & KM Research Group
University of Mannheim,
Germany
heiner@informatik.uni-mannheim.de

ABSTRACT

The "wisdom of crowds" is accomplishing tasks that are cumbersome for individuals yet cannot be fully automated by means of specialized computer algorithms. One such task is the construction of thesauri and other types of concept hierarchies. Human expert feedback on the relatedness and relative generality of terms, however, can be aggregated to dynamically construct evolving concept hierarchies. The InPhO (Indiana Philosophy Ontology) project bootstraps feedback from volunteer users unskilled in ontology design into a precise representation of a specific domain. The approach combines statistical text processing methods with expert feedback and logic programming to create a dynamic semantic representation of the discipline of philosophy. In this paper, we show that results of comparable quality can be achieved by leveraging the workforce of crowdsourcing services such as the Amazon Mechanical Turk (AMT). In an extensive empirical study, we compare the feedback obtained from AMT's workers with that from the InPhO volunteer users providing an insight into qualitative differences of the two groups. Furthermore, we present a set of strategies for assessing the quality of different users when gold standards are missing. We finally use these methods to construct a concept hierarchy based on the feedback acquired from AMT workers.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; H3.1 [**Content Analysis and Indexing**]: Thesauruses

JCDL'10, June 21–25, 2010, Gold Coast, Queensland, Australia.

General Terms

Human Factors, Experimentation

Keywords

Crowdsourcing, Thesaurus learning, Similarity

1. INTRODUCTION

Creating and maintaining concept hierarchies is a costly and cumbersome task. It is normally performed by a group of specialists skilled in both the domain of interest and the relevant methods of formal knowledge representation. As such experts are expensive and in short supply, the discovery of alternative methods of creating and maintaining concept hierarchies would be a major benefit. Significant work has been performed on the automatic creation of classification hierarchies from texts, but these methods often fail to correctly capture semantic relations between topics. In particular, automatic methods are often weak on the task of determining the type of relation that holds between two terms.

These problems have inspired researchers to search for alternative sources of information to augment statistical co-occurrence information in the construction and validation of concept hierarchies. One approach routinely solicits small amounts of information from a concept hierarchies' users while they are engaged in the process of using and maintaining it. An example of such an approach is the InPhO system [11]. This system is based on the involvement of the user community that consists of a relatively small number of domain experts whose expertise is gathered and combined to dynamically generate a taxonomy of philosophical ideas. The general applicability of this approach, however, is hampered by the fact that it relies on the existence and commitment of expert volunteer users.

The objective of this paper is to investigate whether similar results can be achieved in a setting where no group of experts is available and instead a much larger number of non-experts provide the input. This approach, often referred to as "wisdom of crowds", has become very popular recently in the context of so-called Web 2.0 applications and it has been demonstrated that good results can be achieved for

tasks such as annotating unlabeled images [20]. The main question of this paper can be expressed as follows: Is it possible to use the wisdom of crowds to create high quality concept hierarchies in a relatively challenging, abstract domain like philosophy? The contributions of this work are the following:

- We design an experiment for acquiring classification hierarchies from arbitrary web users using Amazon Mechanical Turk (AMT).
- We compare the results provided by non-experts with the results reported in [13].
- We propose effective methods for filtering non-expert feedback based on quality-diagnosing questions.
- We show that the "wisdom of the crowd" performs well when applied in the right way.

The paper is structured as follows. In the next two sections, we discuss the wisdom of the crowds principle, its strengths and weaknesses as well as existing experiences. We then briefly recall the approach used for acquiring a concept hierarchy as described in [11] and present the experimental setup we developed for this work. We close with a discussion of the experimental results and elaborate on lessons learned and potential future work.

2. THE WISDOM OF CROWDS

"Far-flung genius", "distributed intelligence", "innovation communities" - these are exemplary descriptions of the phenomena best characterized as "wisdom of the crowd" [3]. The wisdom of crowds thesis believes in the supposition that many together make better decisions than one person alone. Of course, a concept that combines such far-ranging social implications (see for example Tocqueville's "Tyranny of the majority" [19]) with a high potential also evokes doubts, concerns and critique regarding its validity. These doubts find their expression in the equally used terms "crowd stupidity" or "irrational mobs". Internet guru Jaron Lanier even writes of "digital maoism" in an *Edge* article concerning the Wikipedia encyclopedia [10]. In fact, with the research on collaborative Web 2.0 applications like Wikipedia, Facebook, Del.icio.us and Flickr – supposed applications of farflung genius – two main insights have emerged. First, the strength of such approaches lies in their ability to obtain large amounts of data from volunteer users and to semantically link data obtained from many heterogeneous sources. Second, the organization, validation, and integration of the collected data falls well short of professional standards regarding structural depth and reasoning capabilities. For instance, users tend to assign ambiguous and sometimes entirely meaningless labels to web resources that cannot easily be used in the formalized context of concept hierarchies [18]. In any case, the concept of collaborative tagging is not directly transferrable to our scenario because of a substantial difference: the users of tagging platforms want to see an immediate impact from their work. Very few individuals are willing to tag without even this minimal form of compensation, whether it be for their own information management purposes or some social acknowledgment from a user community. The task in question here is not likely to offer such compensation. Furthermore, users tag the things *they* like to tag. The majority of internet users only linger at places that offer interesting and entertaining content, or pledge clear returns on time and/or cognitive investments. Without providing additional guidance or incentives to users, we cannot rely upon them to provide information where it is most needed. Hence, if we want to use the wisdom of a large enough crowd and not a group of experts or motivated individuals to accomplish our task, we have to attract the crowd's attention, and solicit their help where it is most needed. We can achieve this by using a conceivably easy measure: by paying for targeted feedback.

Even Lanier admits the benefit of the crowd, namely if it is shaped by a tight frame of simple tasks that do not require any coordination or decision making between the working individuals. In "Crowdsourcing" solutions like *InnoCentive*[1], *CrowdFlower*[2], *Wilogo*[3] or *fellowforce*[4] a distributed problem-solving model is implemented that harnesses the creative solutions of a distributed network of individuals, uses an open call format for this, and thus "blends best aspects of open source philosophy and benefits of global business" [3]. Moreover, these solutions mostly focus on specialized tasks such as logo-creation or at least place considerable restrictions on possible task creation. Thus, despite these examples, much more research is needed before crowdsourcing can be viewed as a viable approach in a variety of information extraction tasks. Luckily, a number of platforms have emerged to provide a framework for the creating and experimenting with the crowdsourcing of a variety of general purpose tasks. Probably the best known is the *Amazon Mechanical Turk*[5]. With AMT, Amazon offers extensive options for creating customized questionnaires. The results can be easily processed as they are made available in standard formats. Due to its relatively high publicity (roughly 250,000 tasks available at the time of this writing), it attracts a lot of users and consequently seems most suitable for our scenario.

3. RELATED WORK

The suitability of crowdsourcing in general, as well as paid services such as AMT in particular, has been evaluated for various tasks. Similar to our scenario, Snow et al. [16] evaluated AMT for natural language tasks, including word similarity and word sense disambiguation. They conclude that by means of redundancy, an expert-quality feedback can be achieved and that for this purpose on average 4 non-expert answers are needed, a result confirmed by the experiments in this paper. Sheng et al. [15] examined the effect of redundancy on quality in the context of data mining. To achieve good results, the tasks given to the Mechanical Turk have to be chosen carefully. Effects of task format on feedback quality were evaluated by Kittur et al. [9] who state that validation tests are useful to filter suspicious answers. They also give further advice for the design of the tasks. According to Hsueh et al. [8], the quality of crowdsourcing results for sentiment classification can be improved by eliminating noisy annotators and ambiguous examples. The authors demonstrate that the quality measures in this con-

[1]http://www.innocentive.com
[2]http://crowdflower.com
[3]http://en.wilogo.com
[4]http://www.fellowforce.com
[5]http://www.mturk.com

text are useful for selecting annotations that also lead to more accurate classification models. Alonso et al. [1] describe in detail, how AMT can be used to evaluate the relevance of information retrieval systems. Sorokin and Forsyth [17] used the Mechanical Turk to label images, in particular body parts and shapes on photographs containing people. They experimented with different kinds of annotation tools and described the differences in the results. An alternative approach to attracting a community and getting it to complete the desired task is presented by von Ahn et al. [20, 21]. In this case, users play games, and, in the process of playing, label images, locate labeled objects on images, or gather common-sense knowledge.

4. COLLABORATIVELY EXTENDING AND LEARNING CONCEPT HIERARCHIES

Ontology-based approaches to data organization and integration in specialized domains have produced significant successes, especially within bio- and medical informatics projects (such as the Gene Ontology) and in business applications. One of the major challenges hindering such approaches from successful application to the Web at large is the so-called "knowledge acquisition bottleneck," that is, the large amount of time and money needed to develop and maintain the formal ontologies. In addition, once elaborate and precise ontologies have been created they often lack users to maintain them over time. Since expertise in both ontology design and the relevant domain are required to populate and maintain ontologies, semantic web projects have faced the dilemma of either hiring expensive "double experts" highly-skilled in both ontology design and the relevant domain or face inevitable data and user sparseness [5]. Several researchers have therefore begun to realize that the strength of both approaches could potentially provide a remedy for some of the mentioned weaknesses [4, 2, 6] and have called for the development of the "social-semantic" web, which would combine the social web's ability to obtain large amounts of data from volunteer users with the semantic web's complex and interoperable data representations.

In previous work [11, 12] we presented the InPhO project as one of the first to maintain a dynamically growing knowledge representation of the discipline of philosophy. The system is primarily developed to create and maintain a formal ontology for a well-established, open-access reference work, the Stanford Encyclopedia of Philosophy (SEP). Three features of the SEP make it an ideal environment for developing and testing digital tools to learn and manage ontologies. First, it is substantial and complex: over 1150 entries (>14 million words) of sophisticated humanities content that is beyond the comprehension of any one individual. Second, the SEP is dynamic: new and revised entries come online each month. Finally, it is expert-driven: more than 1,400 professional philosophers serve as its editors and authors. The feedback provided by SEP authors about their areas of expertise provides an overlapping mosaic of expert knowledge. We believe that many online reference works are well-positioned to address the mentioned challenges by making use of their most valuable informational resource: the domain experts who serve as their editors and contributors. With care, expert feedback can be obtained to "approve" the recommendations of automated methods without presuming knowledge of ontology design or placing undue demands on the contributors' time. This feedback can give the modeler a window on the domain which is likely to be the most objective source of knowledge available. The InPhO project successfully maintains a dynamically growing taxonomy of philosophical ideas by leveraging feedback facts provided by a user community consisting of users ranging from interested amateurs to domain experts.

The concepts in the InPhO taxonomy are related over *is-a* relations. Each of these concepts (e.g., *rationalism*) is referred to by a term in InPhO's controlled vocabulary. The problem of determining hierarchical relationships between concepts can be reduced to that of finding hierarchical relationships between terms, that is, extracting hypernym and hyponym relations from text. There are two necessary conditions for a term t_1 to be a hypernym of term t_2: it has to be (a) semantically similar to t_2 and (b) more general than t_2 in the context of the subject area the terms are used in. Conversely, for a term t_1 to be a hyponym of term t_2 it has to be (a) semantically similar to t_2 and (b) more specific than t_2. A large number of measures for the semantic similarity between terms exist. Such measures of similarity and generality have been combined to provide, for any given term, a ranking of possible hyponyms and hypernyms, respectively [11]. The ranking is then presented to InPhO's users to approve or falsify the estimates of semantic relatedness and relative generality of pairs of terms. The relatedness is scored on a five-point scale from highly related to unrelated, and the generality question has four options: same level of generality, *idea*1 is more general than *idea*2, *idea*1 is more specific than *idea*2, and the two are incomparable. The generality of two ideas is deemed incomparable if they are entirely unrelated or if one idea can be both more *and* less general than the other, depending on the context. In this manner, expert feedback can be obtained to "confirm" or "disconfirm" hypotheses about semantic relationships between terms without presuming any knowledge of ontology design.

This use of expert feedback, however, raises three further challenges. For one, while expert feedback may be the highest quality feedback available for the domain, it is hardly infallible, and experts will often be biased in predictable ways, for example by privileging their own preferred area of specialty over others. We have tried to finesse the issue of expert bias by collecting redundant feedback from multiple experts and by looking for inconsistencies, either direct (e.g. expert 1 says A is more general than B whereas expert 2 says B is more general than A) or implied (e.g. inferred through the transitivity of taxonomic relations). Secondly, the presence of inconsistencies raises a further challenge of finding rational strategies to cope with these forms of expert disagreement, preferrably in a way which mitigates expert bias. Thirdly, feedback is collected asynchronously, either as volunteers evaluate pairs or as feedback is solicited during routine tasks, such as during the process of adding and updating encyclopedia articles. To address all of these issues, we have recommended a *dynamic* approach to ontology population and design, on which ontologies are built and populated continuously as feedback is received using a non-monotonic answer set program with stable model semantics. Expert feedback is translated into first-order facts as they come in, and our answer set program is run on these facts daily to flexibly re-construct the global populated ontology.

On this scheme, we use several methods to address the

problem of inconsistent expert feedback[12]. First, each user self-reports a level of expertise (1=amateur, 2=undergrad course, 3=grad course, 4=publication in area) in up to two areas of philosophy. We deal with intra-level and inter-level inconsistencies separately. First, intra-level inconsistencies are settled before feedback facts are passed to the answer program by using a pre-processing "voting" filter, which takes a "majority rules" vote at each strata of expertise. For example, if at expert level 2 we find 4 users asserting that A is more general than B and 2 users asserting that B is more general than A, only the majority opinion at level 2 passes through the filter (where ties are settled by returning to the statistical estimates of generality and similarity). This allows us to screen out "eccentric" expert judgments without being paralyzed by inconsistencies. A further challenge, however, is to flexibly and rationally integrate inter-level inconsistency while making good use of the insight that all user feedback is not created equal. In short, when resolving intra-level inconsistencies, we want to be able to privilege expertise without throwing away possibly useful information contained in responses provided by non-experts. Our current solution to this problem involves a second round of filtering within the answer set program. Candidate taxonomic facts are asserted in the final ontology only when there is evidence for them and no evidence against them, and when two facts (whether directly asserted by users or inferred from user feedback by the answer set program) are inconsistent, the fact at the *lower* level of expertise is said to have strong evidence against it and is discarded. In addition, trust and reliability scores are automatically computed for each user to evaluate her reliability, and these further sources of provenance information can be used in future inconsistency-resolution schemes. The final inferred ontology is thus a mosaic continuously constructed through the flexible integration and cross-validation of partial and overlapping feedback provided by a number of users of varying levels of expertise. The growing knowledge base can be browsed online[6].

While the InPhO project is fortunate enough to continuously collect feedback facts from its volunteer users, the existence of a motivated user community is an exception. The question we mainly address in this paper is whether the InPhO approach can be applied in scenarios where a user community is absent. Instead of relying on volunteers users can we, for instance, pay AMT workers to acquire feedback facts? And what is the quality of these feedback facts?

5. EXPERIMENTAL SETUP

The objective of our experiments is twofold. First, we want to compare the quality of feedback provided by the InPhO community with the feedback provided by the AMT users. Thus, for the first time, we directly compare the quality of contributions provided by a typical Web 2.0 community of volunteer users with those provided by AMT's workers. Second, considering that many real-world scenarios lack de-facto gold standards such as InPhO's set of expert evaluations, we describe and compare different strategies to filter users according to their feedback quality. We believe that these strategies are not only applicable for extending and populating taxonomies but also in other knowledge management scenarios.

[6]http://inpho.cogs.indiana.edu/taxonomy/

At the time of our experiments, the InPhO[7] system had 114 registered users, 45 of which provided one or more of the 4,883 feedback facts. Table 1 shows, for each $1 \leq i \leq 5$, the number of pairs that were evaluated by at least i different users. Among the 114 users, 43 reported the highest level of expertise, meaning that they had published in their respective area, and 45 had finished a graduate class in philosophy. Based on the existing evaluations from the InPhO community we created the dataset of pairs that were given to the AMT workers for evaluation. As a significant overlap with InPhO's users is needed to compare the results, we selected only concept pairs that were evaluated by at least 2 distinct InPhO users, resulting in 1,154 pairs of concepts.

The experience we gained from preliminary small-scale experiments indicated that a rigorous evaluation of the results is impossible if most AMT users evaluate only a small number of concept pairs. To avoid this data sparseness problem, we created single tasks – refered to as Human Intelligence Tasks, HITs – that consisted of 12 distinct pairs of philosophical concepts. This way we obtained at least 12 different evaluations from each AMT user.

For each set of 12 concept pairs we created 5 HITs to obtain at least 5 evaluations by 5 different workers for each distinct pair of concepts. This resulted in 8,640 pairs that were presented to AMT workers in 720 distinct HITs each consisting of 12 concept pairs. A finished HIT was awarded 0.16 USD and the maximum work time for each HIT was set to 20 minutes. The HITs were presented to the AMT workers in the same form as they are presented to InPhO users ensuring equivalent conditions and comparability of the results. Figure 1 depicts the AMT interface with the concept pair "virtue epistemology" and "epistemology."

5.1 Measuring Agreement

There is a large set of available measures to assess the deviation of two statistical variables. In our setting, we are interested in quantifying the agreement of groups of users. Therefore, in our experiments, we always compute the degree of deviation between the feedback facts obtained by different sets of users. As described in Niepert et al. [13], this can be used to determine the disagreement between a user and other users in the same group. In the following, we define the evaluation deviation framework in a more general way, so that it may also be used to compute the evaluation deviation of groups of users. Let U and U' be two sets of users, let A and B be two sets of individuals in the ontology, and let L be the set of possible labels that can be assigned to elements in $A \times B$. Let the label distance $dist : L \times L \to \mathbb{R}^+$ be a function that assigns to each pair of labels a non-negative real number. Let $E = \{(a, b, l, u) \mid a \in A, b \in B, l \in L, u \in U\}$ be the set of 4-tuples representing the evaluations of users in set U and correspondingly let $E' = \{(a, b, l', u') \mid a \in A, b \in B, l' \in L, u' \in U'\}$ represent the evaluations of users in set U'. Note that here user evaluations are assignments of labels in L to elements in $A \times B$ by the users in U and U'.

We define the evaluation deviation measure $D : U \to \mathbb{R}^+$ as

$$D(u) = \frac{1}{|N(u)|} \sum_{(a,b,l,u) \in E} \sum_{(a,b,l',u') \in E', u \neq u'} dist(l, l') \quad (1)$$

[7]http://inpho.cogs.indiana.edu

Minimum overlap	$i = 1$	2	3	4	5
Number of Pairs	3,237	1,154	370	187	92

Table 1: The number of distinct pairs that were evaluated by at least i InPhO users.

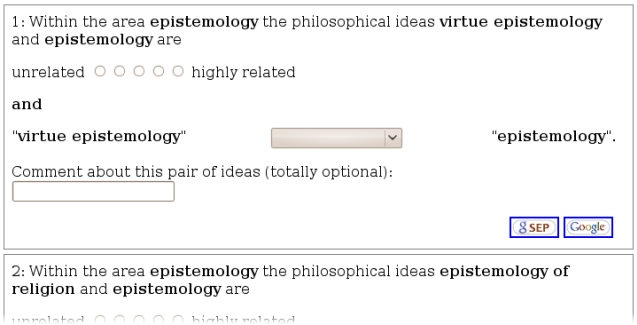


Figure 1: The presentation of a pair in a HIT.

with overlap $N(u) = \{(a, b, l', u') \in E' \mid \exists(a, b, l, u) \in E$ with $u' \neq u\}$. To also measure the quality of the evaluations of groups of users we use the mean of the deviations of the individual users in the groups. The *group deviation* is defined as:

$$D(U) = \frac{1}{|U|} \sum_{u \in U} D(u) \tag{2}$$

with U being the group of users to be compared against the reference group U'. We prefer this measure over standard correlation approaches since it more intuitively reflects the relative degree of disagreement among groups of users and since it is more easily adaptable to different distance measures.

5.2 Telling the Good from the Bad

In this section we describe some strategies that support the assessment of a worker's response quality when no or only a small set of gold standard pairs is available. Possible factors influencing the feedback quality are (a) the time a worker has spent on a specific task and (b) the quality of the worker's feedback on a small set of gold standard pairs that are included in each HIT. Each of the presented strategies is evaluated by comparing the *group deviation* between 13 selected InPhO experts and AMT users selected through the application of certain filters.

Working Time. Here, the underlying idea is that the more time a worker spent on average on the tasks the higher the quality of her feedback. We hypothesized that there exist a group of workers who provide quick random responses in order to maximize their monetary gain while risking the potential disapproval of their HITs. To test this hypothesis, we filtered the set of AMT workers according to the average time needed for completing a HIT. For instance, we excluded all workers who spent less than 100 seconds on their HIT and compared the *group deviation* with the expert group before and after we applied this filter.

Hidden Gold Standard. We placed a small set of diagnostic pairs in each HIT and used the worker's performance on those pairs to assess the quality of the worker's responses.

Minimum HITs	80	30	10	5	2
Number of users	3	6	13	24	41

Table 2: The number of Mechanical Turk users who completed at least a certain number of HITs.

To ensure comparability between all obtained responses, we decided to include the same four concept pairs in every HIT. This means that users who answered more than one set encountered these pairs repeatedly in each set. To disguise this we inserted the pairs in each set at random positions. The chosen concept pairs and the corresponding correct responses are:

Social Epistemology - Epistemology (P_1): Related concepts; social epistemology is more specific than epistemology.

Computer Ethics - Ethics (P_2): Related concepts; computer ethics is more specific than ethics.

Chinese Room Argument - Chinese Philosophy (P_3): Unrelated concepts.

Dualism - Philosophy of Mind (P_4): Strongly related; dualism is more specific than philosophy of mind within the area philosophy of mind.

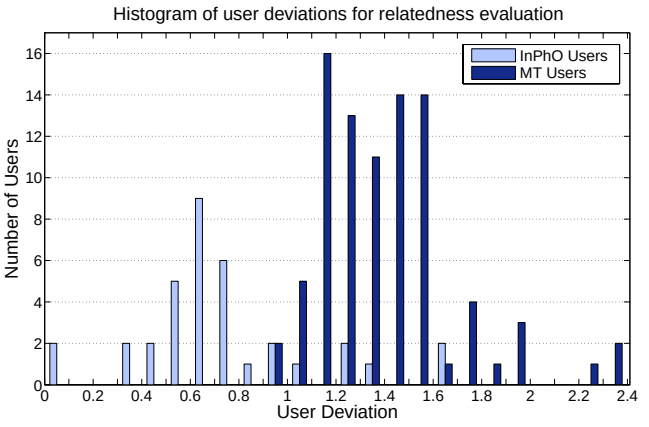
The rationale behind choosing these specific pairs is that the first two pairs should be answerable by everyone without any knowledge about philosophy, only using common sense. The idea is that workers who get one or both of these questions wrong are likely unreliable. The third concept pair presents a more challenging task as some degree of philosophical knowledge is necessary to correctly evaluate this pair. In addition, this is an example where superficial lexical parsing (both concepts contain the term "Chinese") will lead to an incorrect conclusion. Whereas the first two concept pairs evaluate the "lexical performance" of a worker, the third concept pair presupposes "semantic knowledge." The same is true for the fourth pair. Both concepts are highly related, though the relative generality between these two terms is not obvious. Since we ask users to evaluate the pair *relative to the philosophical area philosophy of mind*, the correct response is that dualism is more specific than philosophy of mind.

6. EXPERIMENTAL RESULTS

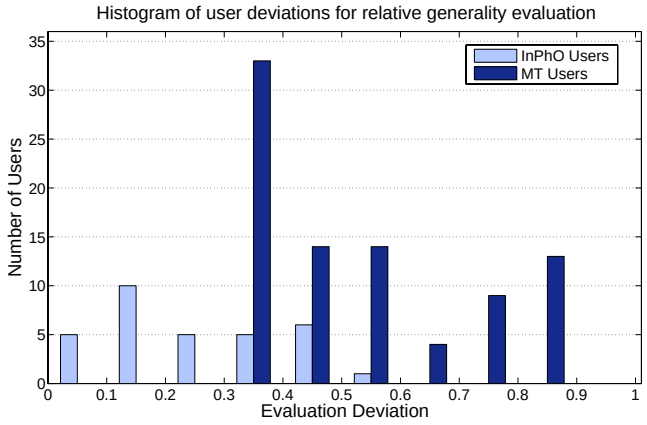
The HITs were completed in 19.7 hours. The average time that a Mechanical Turk user needed for one HIT (12 pairs) was 178 seconds. This means an average hourly rate of 3.25 US$. There were 87 distinct users that completed on average 8.3 HITs. Table 2 shows the distribution, how many users completed at least a given number of HITs. There was only one user who completed all 144 HITs.

6.1 Measuring Agreement

In our experimental setup, we required the users to evaluate a given pair regarding two different aspects: The relatedness of the terms (from unrelated to strongly related) and their relative generality (more specific than, more general than, same generality, incomparable/either).



(a) Relatedness evaluation



(b) Relative generality evaluation

Figure 2: Histogram of inter-group deviations among InPhO and Mechanical Turk users, respectively.

For the labels used to describe the relatedness we define the distance function as

$$dist(l, l') = |l - l'|,$$

where the labels l, l' range from 0 (unrelated) to 4 (strongly related). For stability reasons, we only calculated the deviation for users with an overlap size $|N(u)| \geq 10$. For the relative generality evaluations, we have a set of four independent Labels $L = 0, 1, 2, 3$ with 0="more specific", 1="more general", 2="same generality" and 3="incomparable/either more or less general". For relative generality evaluations, we define the distance function as

$$dist(l, l') = \begin{cases} 0 & l = l' \\ 1 & l \neq l'. \end{cases}$$

Inter-group Agreement. We use Equation 1 with $U = U'$ to compute the inter-group deviation once for the InPhO users and once for the Mechanical Turk users. Figure 2 shows the results for the deviation on the relatedness and relative generality evaluations. Since we required an overlap of at least 10 concept pairs, we compared 35 InPhO users with each other. The number of 87 AMT workers was not reduced, as we ensured in our experimental setup an overlap of every user with at least 4 other users over 12 pairs ($|N(u)| \geq 48$). The result shows that the AMT workers perform significantly worse than the InPhO users regarding their internal agreement on the correct answers for the given pairs. That means that the answers were not as consistent as the answers given by the InPhO community, possibly indicating that the AMT responses are of highly-variable quality.

Comparison with Experts. Measuring the quality of answers is not an easy task, as the relation of terms and the perception of relatedness is very subjective, and even human experts only agree up to a point on the correct answer. In our setting, we have the experts of the InPhO system and we can use their feedback as a de-facto gold standard. We singled out a set of 13 experts, all of whom have published in their area of philosophy, and used this set as the gold standard for all subsequent evaluations. Figure 3 shows the histogram of evaluation deviations, this time with the experts forming the reference set. Of course, these expert users were removed from the InPhO users set

It can be seen that the deviation from the experts' answers, and thus the quality of the answers, is much more variable for the Mechanical Turk users than for the InPhO users. This is not surprising, as the InPhO users performed already quite well in the internal evaluation, including the experts. It is rather promising, however, that there is a significant number of Mechanical Turk users who perform in the same deviation range as the InPhO users. Thus, there is hope that a larger set of the Mechanical Turk users would provide high quality responses.

The results for the relative generality (3(b)) look even better. Probably due to the categorical "right or wrong" definition of the distance, the histogram curve is not as smooth as for relatedness. Instead we have a quite clear distinction between a set of users who performed well and a large set of poorly performing ones, as a deviation of 1 means there is complete disagreement with all the expert users' feedback facts.

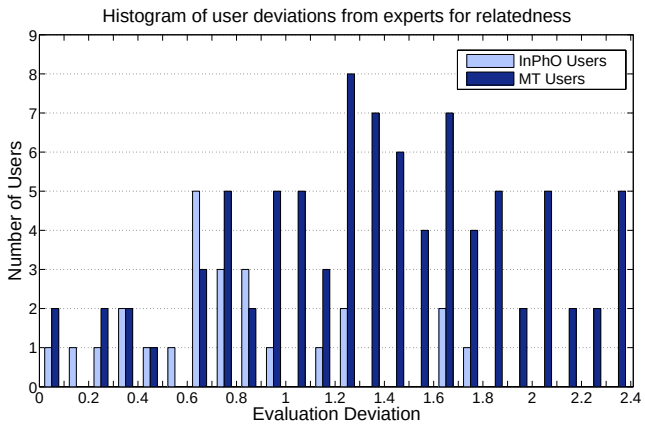
6.2 Telling the Good from the Bad

In the following, we test the two proposed strategies in an attempt to distinguish between reliable and unreliable AMT workers.

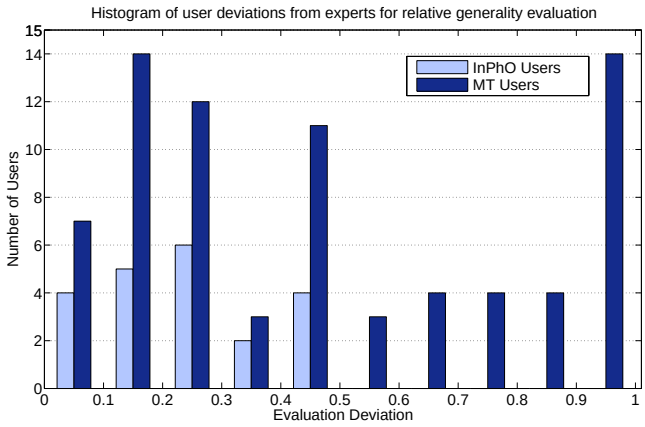
Working Time. The first approach is based on the hypothesis that unreliable users take less time to think about their answers. Thus we try to filter out users based on their average completion time for a single HIT. Figure 4 shows the impact of this filter on the number of excluded workers and the resulting variations in group deviation values compared to the expert reference set.

The results clearly indicate that the completion time is not a good feature for assessing user feedback quality. Both relatedness and relative generality (Figure 4(b)) stay roughly at the same level. The graph also demonstrates that the quality of the responses for relatedness and relative generality are strongly correlated (Kendall τ : 0.36, Spearman ρ: 0.52).

Hidden Gold Standard. The most straight-forward way to distinguish between reliable and unreliable performers is by



(a) Relatedness evaluation



(b) Relative generality evaluation

Figure 3: Histogram of deviations of InPhO and Mechanical Turk users, respectively, compared to expert group.

comparing the workers' responses to a set of gold standard concept pairs for which we know the relatedness and relative generality. To allow for this test, we included four concept pairs ($P_1 \dots P_4$) into each HIT, as explained in Section 5. For these pairs, there exist correct answers on which all InPhO experts agreed. We considered the following answers as correct:

P_1**:** Relatedness ≥ 3 (the two highest levels of relatedness) and relation "more specific than".

P_3**:** Relatedness ≥ 3 and relation "more specific than".

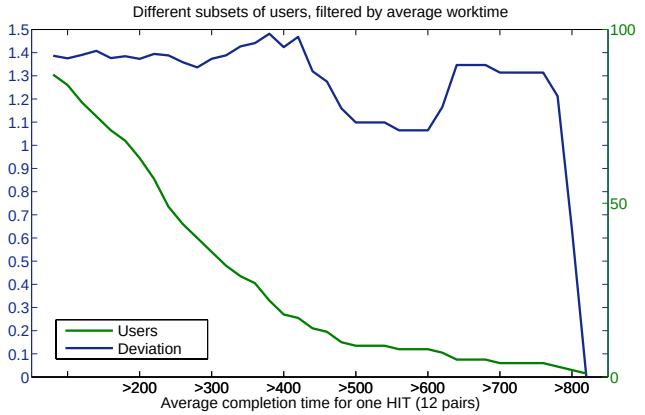
P_1**:** Relatedness $= 0$ (unrelated).

P_4**:** Relatedness ≥ 3 and relation "more specific than".

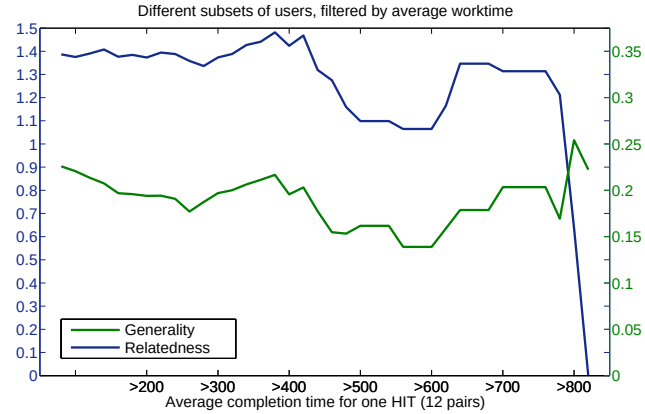
Table 3 lists the number of users who evaluated the given pair correctly, as well as the overall number of users who answered it. We received several answers for the AMT workers who completed more than one HIT as the pairs were repeatedly included. To maintain comparability, we only used the worker's response for the first HIT. It is notable that the InPhO community seems to have more problems with the "easy" pairs P_1 and P_2 than the AMT workers. This is probably due to the low number of cases. The Mechanical Turk

	InPhO Users	MT Users
P_1	7/10 (0.70)	52/87 (0.60)
P_2	2/3 (0.67)	50/87 (0.57)
P_3	2/2 (1.00)	20/87 (0.23)
P_4	5/6 (0.83)	32/87 (0.37)

Table 3: Number of users who answered the pairs correctly.



(a) Number of users and the resulting deviation for the relatedness evaluation.



(b) Deviation results for relatedness and relative generality evaluation.

Figure 4: Effects, if all users are filtered who used less time on average to complete a single HIT (12 pairs)

users perform best for these pairs with roughly 60% of them providing the correct responses.

AMT workers had the most problems with the pair P_3 ("Chinese Room Argument" - "Chinese Philosophy"), but performed better on the evaluation of P_4 ("Dualism" - "Philosophy of Mind"). To get a better understanding of the dependencies between the four pairs of questions, we calculated the conditional probabilities for the correctness of a pair, given that another pair was answered correctly (Table 4). Interestingly, there is a high probability (around 80%) that, if P_1 is answered correctly then P_2 is also answered correctly and vice versa. As both pairs can be answered correctly by just using some common "lexical" sense, we con-

	P_1	P_2	P_3	P_4
P_1	1.00	0.77	0.21	0.50
P_2	0.80	1.00	0.22	0.56
P_3	0.55	0.55	1.00	0.40
P_4	0.81	0.86	0.25	1.00

Table 4: Probability for an MT user to answer the pair in the column correctly, if the pair in the row is correctly answered.

(a) Relatedness evaluation

Filter	Users	$D(U)$	Range $D(u)$
$P_1 \wedge P_2 \wedge P_3 \wedge P_4$	7	0.60	0.00 – 1.00
$P_1 \wedge P_2 \wedge P_3$	10	0.87	0.00 – 1.78
$P_1 \wedge P_2 \wedge P_4$	23	0.84	0.00 – 1.41
$P_1 \wedge P_2$	40	1.11	0.00 – 1.96
All MT users	87	1.39	0.00 – 2.96
InPhO users	25	0.77	0.00 – 1.75
Random	—	1.80	—

(b) Relative generality evaluation

Filter	Users	$D(U)$	Range $D(u)$
$P_1 \wedge P_2 \wedge P_3 \wedge P_4$	7(5)	0.12	0.00 – 0.22
$P_1 \wedge P_2 \wedge P_3$	10(8)	0.14	0.00 – 0.27
$P_1 \wedge P_2 \wedge P_4$	23(20)	0.15	0.00 – 0.45
$P_1 \wedge P_2$	40(35)	0.21	0.00 – 0.59
All MT users	87(78)	0.45	0.00 – 1.00
InPhO users	21	0.23	0.00 – 0.47
Random	—	0.75	—

Table 5: Effect of different filters on the set of Mechanical Turk users

sider their correct evaluation as a minimum requirement that a user has to fulfill. The probabilities for the hardest pair P_3 are surprising, as answering it correctly does not seem to be a good indicator for the correct response on other pairs (about 50% for each). A better predictive property has pair P_4 as workers who answered it correctly also answered P_1 and P_2 correctly with a probability of over 80%. Using these findings, we defined several configurations to filter the users, based on their answer on $P_1 \dots P_4$. Table 5 summarizes the results of these experiments. The filter criterion is defined in a Boolean way, with P_i indicating that the response for P_i has to be correct for the worker to pass the filter. We compared the resulting groups both with the performance of the InPhO community and with the performance of a user who responds at random.

Note that there is no evaluation on the relative generality if a user rates a pair as unrelated. Thus the number of users for which a deviation can actually be computed is reduced and given in parentheses. The results show that, with the most restrictive filter setting, it is possible to achieve a higher agreement with the experts than the InPhO community. Of course, this comes at the price of sacrificing a lot of the evaluations. Asking for this level of quality would require many more completed HITs to collect the needed number of responses. The simple filter $P_1 \wedge P_2$ significantly improves the quality of the results compared to the whole set of MT users. It is still worse than the InPhO community for the relatedness evaluation but outperforms it slightly for the relative generality (0.21 compared to 0.23). Adding P_3 confirms our hypothesis, based on the conditional probabilities, that the users had problems with this pair. However, evaluating it correctly does not imply a generally high response quality. With this filter the number of users is reduced to only 10. The result for the filter $P_1 \wedge P_2 \wedge P_4$ shows that this configuration performs even better, while leaving a much bigger set of 23 users.

6.3 Financial Considerations

Using AMT within an approach means obviously that money is involved. Thus, for a full evaluation of the results, we not only have to focus on the feedback quality we can reach but also the financial price we have to pay for it. Table 6 lists some figures that illustrates the relationship between different filter settings and the number of obtained concept pairs.

Filter	Pairs	Evaluations	C_{Pair}	C_{Eval}
—	1138	5690	0.111	0.022
$P_1 \wedge P_2$	1074	1909	0.117	0.066
$P_1 \wedge P_2 \wedge P_3$	215	215	0.586	0.586
$P_1 \wedge P_2 \wedge P_4$	1018	1558	0.124	0.081
$P_1 \wedge P_2 \wedge P_3 \wedge P_4$	183	183	0.689	0.689

Table 6: The number of unique pairs and single evaluations we gather from different sets of users, as well as the costs in US-Dollar per pair and per evaluation.

For our whole experiment we paid 126 US-Dollar. This is 0.11 USD per concept pair and 0.02 USD per evaluation. With the lowest filtering we still obtained 1,074 pairs barely increasing the price to 0.12 USD per pair. However, the number of usable evaluations was reduced to 1,909. If we would like to have the same amount of redundancy that the experiment was designed for, we would had have to pay about 376 USD. For the highest quality of feedback ($P_1 \wedge P_2 \wedge P_3 \wedge P_4$), the costs for 1,138 pairs are estimated at 784 USD, for 5,690 evaluations we estimate 3,920 USD. Of course, these sums are only estimations, based on the assumption that the coverage of pairs would scale across our whole set of pairs with a proportional increase of HITs.

6.4 Constructing the Concept Hierarchy

To apply our answer set program [12] to the data gathered from the AMT workers, we have to determine, for each of the workers, an expertise level between 0 (no expertise) and 3 (high expertise). The answer set program we have developed for this task considers these expert levels when resolving conflicting feedback facts, as described in section 4. We decided to exclude all workers who evaluated all of the gold standard questions incorrectly.

Then, we again used the filter configurations described above to determine the expertise level:

- Users who answered all test pairs correctly ($P_1 \wedge P_2 \wedge P_3 \wedge P_4$) were considered as level 3.
- Users who answered pair 1, 2 and 4 ($P_1 \wedge P_2 \wedge P_4$) correctly were considered as level 2.

- Users who answered only pair 1 and 2 ($P_1 \wedge P_2$) correctly were considered as level 1.

The resulting concept hierarchy can be browsed online at http://inpho.cogs.indiana.edu/amt_taxonomy/.

6.5 Ethical Considerations

After considering the statistical results and hard financial facts, we should remember that the Mechanical Turk is no computer, no algorithm or approach like many others developed in computer science. Even if it provides an API that allows a seamless integration into computer systems, the actual work is done by real human beings.

Online piece work like AMT has been criticized as possibly leading to "digital sweatshops," in which the inexpensive labor of citizens from developing countries is exploited to complete menial tasks that others are unwilling to do [22]. While it is beyond the scope of this paper to provide a detailed analysis of the social or ethical implications of the use of services like Mechanical Turk, a few preliminary comments are in order. The ethical factors involved in AMT use for such a project can be organized into two groups– user-level considerations (e.g. pertaining to the well-being of workers) and systemic considerations (e.g. whether AMT itself encourages unjust or unethical practices). We discuss each in turn, with an eye towards practical advice for other projects.

Let us first consider the welfare of the workers completing the HITs. First, it is not clear that the demographics of Turkers supports the digital sweatshop narrative. While HIT providers are forbidden by AMT terms of service from asking demographic questions, a study conducted by Ross et al. has found that while a minority of Turkers are citizens of developing nations and an even smaller minority depend upon AMT for a significant portion of their income, a significant majority of Turkers reside in the U.S. and many have relatively high household incomes [14]. Many users reported that they complete AMT tasks as a diversion, suggesting that the tasks themselves are not as onerous as one might suppose. Secondly, for the minority of users who *do* live in developing countries and depend on AMT for primary income, one might compare the wages and conditions of AMT tasks to other employment opportunities locally available to these users. No forms of coercion other than payment are directed towards the users, and users have significant freedom concerning work hours and conditions. A study conducted by Horton [7] found that AMT users reported finding AMT employers as fair as or more fair than local employers (though there are serious issues with the sample in this study, given that AMT was itself used to conduct the experiment). One persistent worry related to worker exploitation has to do with the fact that employers can, at their discretion, opt to reject HITs and not pay users (while possibly still making use of the data), and users have no ability to appeal this decision. However, users can see HIT-provider rejection rates before accepting a task, and web sites have sprung up to evaluate HIT-providers – so AMT users may peruse reviews of tasks completed by other users before choosing to participate. Where applicable, we recommend that employers warn users that some form of quality-control will be used to evaluate HIT responses before compensation will be provided (though providing specific information about the controls would of course erode their utility).

Another more systemic concern is that users have little or no inkling as to the ends to which their labor will be put, and some have used AMT for nefarious ends such as writing fraudulent product reviews. While much of this should be settled by better filtering of HITs by Amazon, we recommend that employers give users some idea as to the ends to which their responses will be put–though employers must often toe a fine line, familiar to psychologists, between enabling informed consent and biasing subjects.

7. CONCLUSION

In this paper, we examined the prospects of a paid crowdsourcing service, namely Amazon Mechanical Turk, to complement the role of a community project in the context of learning and populating a taxonomy for the discipline of philosophy. The experiments are promising but several important aspects have to be taken into account. Generally, the answers of AMT users are of varying quality, particularly if they are directly compared to a community of experts and motivated laypersons. But the comparison also revealed that there are AMT users who perform very well, and simple filtering rules can sometimes be used to indentify them. In line with findings of other authors [1, 8, 9, 16], we showed that it is possible to achieve high quality results, even outperforming the community. In particular, we accomplished this with the following steps:

- Every pair was evaluated 5 times by 5 different users to ensure the necessary redundancy;
- We included a small set of concept pairs for which we could objectively determine a correct answer; and
- Based on the responses to our test pairs, we filtered the users to improve the overall quality of the answers.

With these steps and a moderate filtering policy we achieved a feedback quality comparable to that of the InPhO community. The remaining high quality users still covered $1,018$ of the original $1,138$ (89%) concept pairs that we collected for our experiments. The completion of the AMT task cost less than 1 day and totaled 126 US-Dollars. Based on the workers' feedback we were able to create a concept hierarchy which can be browsed online at
http://inpho.cogs.indiana.edu/amt_taxonomy/.

An important property of the method presented in this paper is that it does not rely on any existing data, gold standards or training data provided by experts. The next steps include the further refinement of the thesaurus learning process using Amazon Mechanical Turk. Furthermore, we intend to transfer the approach from the domain of philosophy to other domains. Another promising avenue of future work is the employment of more sophisticated algorithm such as support vector machines to classify AMT's workers according to their feedback quality.

Acknowledgments

During the preparation of this manuscript, Cameron Buckner and Colin Allen were supported by an award from the National Endowment for the Humanities Division of Preservation and Access.

8. REFERENCES

[1] Omar Alonso, Daniel E. Rose, and Benjamin Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[2] Anupriya Ankolekar, Markus Krötzsch, Duc Thanh Tran, and Denny Vrandecic. The two cultures: Mashing up web 2.0 and the semantic web. *Journal of Web Semantics*, 6(1):70–75, 2008.

[3] Daren C. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *The International Journal of Research into New Media Technologies*, 14 (1):75–90, 2008.

[4] John G. Breslin, Alexandre Passant, and Stefan Decker. *The Social Semantic Web*. Springer-Verlag, 2010.

[5] Cameron Buckner, Mathias Niepert, and Colin Allen. From encyclopedia to ontology: Toward dynamic representation of the discipline of philosophy. *Synthese*, 2010 (forthcoming).

[6] Tom Gruber. Collective knowledge systems: Where the social web meets the semantic web. *Journal of Web Semantics*, 6(1):4–13, 2008.

[7] John Horton. The condition of the turking class: Are online employers fair and honest? *CoRR*, abs/1001.1172, 2010.

[8] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. Data quality from crowdsourcing: a study of annotation selection criteria. In *HLT '09: Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Morristown, NJ, USA, 2009. Association for Computational Linguistics.

[9] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *CHI '08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 453–456, New York, NY, USA, 2008. ACM.

[10] Jaron Lanier. Digital maoism: The hazards of the new online collectivism. *Edge*, 183, March 30th 2006.

[11] Mathias Niepert, Cameron Buckner, and Colin Allen. A dynamic ontology for a dynamic reference work. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 288–297, New York, NY, USA, 2007. ACM.

[12] Mathias Niepert, Cameron Buckner, and Colin Allen. Answer set programming on expert feedback to populate and extend dynamic ontologies. In *Proceedings of the Twenty-First International Florida Artificial Intelligence Research Society Conference, May 15-17, 2008, Coconut Grove, Florida, USA*, pages 500–505, 2008.

[13] Mathias Niepert, Cameron Buckner, and Colin Allen. Working the crowd: Design principles and early lessons from the social-semantic web. In *Proceedings of the Workshop on Web 3.0: Merging Semantic Web and Social Web at ACM Hypertext*, 2009.

[14] J. Ross, A. Zaldivar, L. Irani, and B. Tomlinson. Who are the turkers? worker demographics in amazon mechanical turk. Technical report, Department of Informatics, University of California, Irvine, USA, Technical Report SocialCode-2009-01, 2009.

[15] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622, New York, NY, USA, 2008. ACM.

[16] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[17] Alexander Sorokin and David Forsyth. Utility data annotation with amazon mechanical turk. *Computer Vision and Pattern Recognition Workshop*, 0:1–8, 2008.

[18] Fabian M. Suchanek, Milan Vojnovic, and Dinan Gunawardena. Social tags: meaning and suggestions. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge mining*, pages 223–232, New York, NY, USA, 2008. ACM.

[19] Alexis de Tocqueville, Henry Reeve, and John C. Spencer. *Democracy in America [microform] / by Alexis de Tocqueville, translated by Henry Reeve. With an original preface and notes by John C. Spencer.* G. Dearborn & Co., New York, 1838.

[20] Luis von Ahn and Laura Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM.

[21] Luis von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: A game for collecting common-sense knowledge. In *ACM Conference on Human Factors in Computing Systems, CHI Notes*, 2006.

[22] Jonathan Zittrain. Work the new digital sweatshops. *Newsweek*, December 8, 2009.