Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts
AAAI Technical Report CR-13-01

Ontology Quality Assurance with the Crowd

Jonathan M. Mortensen, Mark A. Musen, Natalya F. Noy
Stanford Center for Biomedical Informatics Research
Stanford University, Stanford, CA, 94305-5479

Abstract

The Semantic Web has the potential to change the Web as we know it. However, the community faces a significant challenge in managing, aggregating, and curating the massive amount of data and knowledge. Human computation is only beginning to serve an essential role in the curation of these Web-based data. Ontologies, which facilitate data integration and search, serve as a central component of the Semantic Web, but they are large, complex, and typically require extensive expert curation. Furthermore, ontology-engineering tasks require more knowledge than is required in a typical crowdsourcing-task. We have developed ontology-engineering methods that leverage the crowd. In this work, we describe our general crowdsourcing workflow. We then highlight our work on applying this workflow to ontology verification and quality assurance. In a pilot study, this method approaches expert ability, finding the same errors that experts identified with 86% accuracy in a faster and more scalable fashion. The work provides a general framework with which to develop crowdsourcing methods for the Semantic Web. In addition, it highlights opportunities for future research in human computation and crowdsourcing.

The Semantic Web and Ontology Engineering

The Semantic Web seeks to transform the Web from linked documents to linked data. One challenge in reaching this goal is managing, aggregating, and curating the massive amount of data and knowledge on the Web. Ontologies, which facilitate data integration and search, serve as a central component of the Semantic Web, but they are often large, complex, and typically require extensive expert curation. For instance, in the National Center for Biomedical Ontology's BioPortal, there are over 350 ontologies, with many containing thousands of concepts and millions of relations (Musen et al. 2012). As ontology scale and complexity increase, so does the challenge of engineering them. Such large artifacts inherently contain errors; furthermore, automated methods to identify such domain-specific errors are limited in either performance or scalability. Thus, traditionally, expert curation is necessary to identify domain-specific errors in an ontology. For example, through manual inspection, Rector and colleagues identified significant modeling errors (e.g., an axiom stating: **Foot** *part-of* **Pelvis**) in SNOMED CT, a large, widely-used biomedical ontology (Rector, Brandt, and Schneider 2011). Even so, manual curation cannot scale due to cost and time constraints.

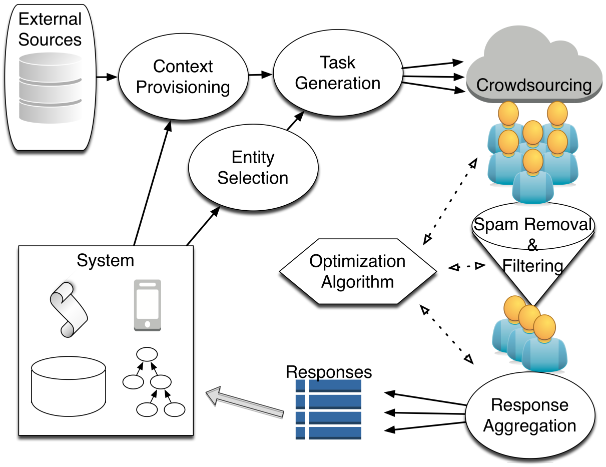

Figure 1: General workflow for crowdsourcing (read clockwise from the upper left).

Recently, we have begun developing a method that uses crowdsourcing to identify domain-specific ontology errors scalably and accurately (Mortensen et al. 2013; Mortensen, Musen, and Noy 2013; Noy et al. 2013). In this work, we describe a general crowdsourcing workflow. In the context of this workflow, we highlight our method to perform ontology quality assurance with the crowd.

A Crowdsourcing Workflow

In developing a crowd-backed method for ontology engineering, we arrived at a general workflow that a typical crowdsourcing method follows. Figure 1 provides an outline of the workflow. A description of each element follows.

Entity and Context Selection A crowdsourcing method begins with the crowdsourcer selecting the entities from a system upon which to perform the task. For example, in ontology alignment, a developer might select concepts from two ontologies to align. In addition, one typically provides a worker with context, which may not reside in the main system, to assist with the task completion. In the alignment example, the developer might provide concept parents and children, or concept synonyms.

Task Generation With entities and context obtained, one then creates a task. This includes determining the type of task, how workers will complete the task, and its general presentation. Continuing with the ontology alignment example, the task might be a binary choice indicating whether concepts are the same or different. Then, the requester would determine the appropriate presentation of the concepts, their context, and user selection boxes.

Crowdsourcing The tasks are then submitted to a crowdsourcing marketplace (e.g., Amazon's MTurk). These marketplaces have various parameters to specify, and after task submission, return a set of responses from workers.

Aggregation and Filtering With the responses from the marketplace, a requester filters them using some criteria. The criteria normally are defined in an effort to remove low quality or spam responses. Finally, one aggregates the remaining responses to a final single response. In ontology alignment, this might be a final alignment decision indicating if the concepts are equivalent or not. Based on logic in an optimization algorithm or "controller", one might continue submitting tasks, repeat the entire process, or end the process.

Crowd-Backed Ontology Verification

Using our general crowdsourcing framework, we developed a method for ontology quality assurance that verifies the correctness of statements in an ontology (Mortensen et al. 2013). For instance, we might want to verify that "**Person** is a kind of **Mammal**". As a pilot study, we applied this method to a set of known correct and incorrect statements in SNOMED CT that Rector et al. previously identified (2011). The following describes this method in terms workflow.

System Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT)

Entities Seven logical statements in SNOMED CT that Rector and colleagues identified as incorrect and seven similar statements that are correct in SNOMED CT. Each statements relates two concepts (e.g., **Heart** *is-a* **Organ**).

Context Concept definitions from the Unified Medical Language System and Wikipedia.

Task We present workers with two concept definitions and the statement in natural language (e.g., **Heart** is a kind of **Organ**). Workers then select True or False.

Crowdsourcing We submit tasks to Amazon Mechanical Turk. We requested 40 workers respond to each of the 14 tasks and pay $0.02/response.

Aggregation & Filtering We perform Bayesian inference with a beta distribution to describe the crowds confidence that a statement is True (Similar to Ipeirotis et al. 2013).

With majority voting (a special case of the aggregation method), the crowd was 86% accurate in differentiating correct versus incorrect statements in SNOMED CT. Of note, on average, any single worker performed 15% less accurately than the aggregate response.

Discussion

The results of the pilot study suggest that the crowd can assist with performing ontology verification at scale. In future, we plan to increase the study size, apply more advanced response aggregation techniques, and investigate methods that balance cost, performance and speed. The success of this method indicates that other Semantic Web tasks may be amenable to crowdsourcing.

This work serves as an example of a deep knowledge task, whose completion requires an expert or individual with a deep understanding of the domain. However, experts are costly, and cannot scale to the size of the entire Web. Indeed, many Semantic Web tasks are of this type (e.g., performing advanced and complex classification of scientific data). Therefore, future research should focus on developing human computation and crowdsourcing methods that accurately complete such tasks in a scalable, fast, and cost-effective manner.

Conclusion

Human Computation can enable management of large amounts of data and will likely become a common tool for Semantic Web researchers. Ontology engineering, where human curation is required in a scalable fashion, is one particular area that crowdsourcing can assist. In this work, we described a general crowdsourcing workflow and an application of this workflow to the task of ontology quality assurance. This work serves as motivation of research opportunities in human computation for tasks that require deep expert knowledge of a domain.

Acknowledgments This work is supported in part by grants GM086587, HG004028, and LM007033 from the NIH

References

Ipeirotis, P. et al. 2013. Repeated labeling using multiple noisy labelers. *Data Mining and Knowledge Discovery* 1–40.

Mortensen, J. M.; Alexander, P. R.; Musen, M. A.; and Noy, N. F. 2013. Crowdsourcing Ontology Verification. In *International Conference on Biomedical Ontology*.

Mortensen, J. M.; Musen, M. A.; and Noy, N. F. 2013. Crowdsourcing the Verification of Relationships in Biomedical Ontologies. In *AMIA Annual Symposium*.

Musen, M. A. et al. 2012. The National Center for Biomedical Ontology. *J. Amer. Med Informatics Assoc.* 19:190–195.

Noy, N. F.; Mortensen, J. M.; Alexander, P. R.; and Musen, M. A. 2013. Mechanical Turk as an Ontology Engineer? Using Microtasks as a Component of an Ontology Engineering Workflow. In *Web Science*.

Rector, A. L.; Brandt, S.; and Schneider, T. 2011. Getting the foot out of the pelvis: modeling problems affecting use of SNOMED CT hierarchies in practical applications. *J. Amer. Med Informatics Assoc.* 18(4):432–440.