ADM Institute Seed Research Funding

Progress Report (4/11-3/12)

Name	Tarek Abdelzaher
Research Project	Empowering the Human: A Crowd-sourcing Perspective
Department/s	Computer Science
Date	5/7/2012

Please provide a brief response below.

---

1. Please provide a brief statement of the project accomplishments thus far.

Below, I briefly summarize the proposed work then present accomplishments on each proposed task. The proposal was to explore the benefits of *crowd-sourcing* as a means to gain insight into complex social and physical situations. The goal, as quoted from the proposal, was "to build an automated information collection and management service that allows a provider to prompt large groups of individuals to share data via their cell-phones on selected issues of concern (in the form of text, voice, or images, depending on cell-phone capabilities). The service back-end then performs data cleaning and analytics to demonstrate the viability of extracting reliable information from the large number of individually potentially-unreliable or biased sources that might join the data collection campaign." The project proposed four activities to demonstrate four important aspects of crowd-sourcing. These proposed activities and progress on each are presented below.

- **Activity #1: *Demonstrate viability of extracting accurate information using crowd-sourcing.*** The project promised to "investigate the degree to which crowd-sourcing data can be cleaned to yield useful information".

  Progress:

  a) Survey: The PI surveyed the state of the art in crowd-sourcing, otherwise known as "human-centric sensing". This survey has recently appeared in a publication of the *Philosophical Transactions of the Royal Society*:

  [1] Mani Srivastava, Tarek Abdelzaher, Boleslaw K. Szymanski, "Human-centric Sensing," *Philosophical Transactions of the Royal Society, special issue on Wireless Sensor Networks*, Vol. 370, No. 1958, pp. 176-197, January 2012.

  b) Zero-prior-knowledge crowd-sourced data cleaning algorithms: A challenge in crowd-sourcing is that the reliability of data sources (e.g., the individual farmers) is often not known to the data collector a priori. Similarly, the observations shared cannot be readily verified. This poses a problem in assessing which observations are true and which are not. Past literature allows estimation of reliability of measurements if reliability of sources is known and vice versa, but when both are unknown, the problem is open. The PI developed new analytic foundations for jointly assessing the reliability of both *data* and *sources*, and for computing confidence intervals in such reliability estimates. The accomplishment is novel is that it computes these estimates *without prior knowledge* of either source reliability or correctness of individual measurements. The work allows extraction of *quantifiably reliable* information from unreliable data collected from sources of unknown reliability (as might be the case

when collecting information on harvest supply chain state from remote volunteers. The new techniques led to two recent publications:

[2] Dong Wang, Hieu Le, Lance Kaplan, Tarek Abdelzaher, "On Truth Discovery in Social Sensing: A Maximum Likelihood Estimation Approach," *In Proc. 11th ACM/IEEE Conference on Information Processing in Sensor Networks (IPSN)*, April 2012. (15% acceptance rate)

[3] Dong Wang, Lance Kaplan, Tarek Abdelzaher, Charu Aggarwal, "On Scalability and Robustness Limitations of Real and Asymptotic Confidence Bounds in Social Sensing," *9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Seoul, Korea, June 2012. (19% acceptance rate)

- **Activity #2: *Demonstrate scalability*.** To illustrate scalability, the project promised to "use large volumes of Twitter data to reliably reconstruct sequences of reported events, extracted from large human populations". The goal was to compare facts extracted from crowd-sourcing to ground truth obtained from international media, thus demonstrating scalability and accurate fact-finding.

  Progress:

  A preliminary investigation was conducted on scalability issues as well as trade-offs between the volume of processed data (e.g., amount of received tweets, which were used in lieu of phone "texts"), data diversity, and the accuracy of facts extracted from these data. These results were published in a recent invited paper:

  [4] Md Yusuf S Uddin, Md Tanvir Al Amin, Hieu Le, Tarek Abdelzaher, Boleslaw Szymanski, Tommy Nguyen, "On Diversifying Source Selection in Social Sensing," In Proc. *9th International Conference on Networked Sensing Systems (INSS),* Antwerp, Belgium, June 2012. (Invited paper)

- **Activity #3: *Demonstrate viability of data fusion from voice, text, and images*.** The project promised to leverage related ARL funding: "Abdelzaher currently leads an ARL-funded project on heterogeneous data fusion from unstructured sources in military applications. Algorithms developed in this project will be re-tooled and tested to demonstrate fusion of heterogeneous crowd-sourcing data sources."

  Progress:

  An algorithm was developed in the aforementioned project to automatically identify correspondence relations between text and images. In particular, the algorithm focused on recognizing the case where images corroborate text (note that, both text and images are common data formats that can be delivered by contemporary phones in a crowd-sourcing scenario). It is fairly straightforward for a human to recognize that a certain image of a flood or forest fire corroborates text reports of the same event. It is much harder for a machine to do it. The contribution of this activity was to enable automated machine-processing of text and image data for identifying instances of mutual corroboration. Results were integrated with the fact-finding algorithms described above to generate an "illustrated story" of transpired events from text and image data collected via crowd-sourcing. A demo of this capability is available on demand.

- **Activity #4: *Incentives*.** The project promised to address "incentives and recruitment issues for adequate representation of target populations". While the PI has not advanced the state of the art in that area, a survey of current techniques was included in publication [1] mentioned above.

As promised in the proposal, the aforementioned proof-of-concept effort helps establish the viability of developing a virtual information highway based on crowd-sourcing. It paves the way for a larger engagement that applies the proposed mechanisms to the problem of understanding and reducing post-harvest loss.

2. Has the project completed its objectives? Yes or No. If no, please describe the activities planned for the remainder of the project (4/12-3/13).

The PI believes that the project has accomplished its basic goals:

(i) It demonstrated the viability of extracting reliable information from generally unreliable sources and noisy data, which is one of the main hurdles in crowd-sourcing.

(ii) It demonstrated scalability of data collection and processing.

(iii) It demonstrated viability of data fusion from text and images.

(iv) It surveyed techniques to offer participation incentives

It remains to put these parts together into an integrated system that enables stakeholders to carry out and incentivize crowd-sourcing campaigns. Note that, the objective of current funding was to develop a proof of concept. With that (almost) done, the PI is open to starting an expanded project, with support of the ADM Institute, that goes beyond proof-of-concept, and has more emphasis on the specifics of reducing post-harvest loss.

---

3. Please attach (or provide a URL to products electronically available) work products that has been completed.

Please refer to the above-mentioned publications.