Noname manuscript No.
(will be inserted by the editor)

Maximizing Benefits from Crowdsourced Data

Geoffrey Barbier · Reza Zafarani · Huiji Gao · Gabriel Fung · Huan Liu

Received: date / Accepted: date

Abstract Crowds of people can solve some problems faster than individuals or small groups. A crowd can also rapidly generate data about circumstances affecting the crowd itself. This crowdsourced data can be leveraged to benefit the crowd by providing information or solutions faster than traditional means. However, the crowdsourced data can hardly be used directly to yield usable information. Intelligently analyzing and processing crowdsourced information can help prepare data to maximize the usable information, thus returning the benefit to the crowd. This article highlights challenges and investigates opportunities associated with mining crowdsourced data to yield useful information, as well as details how crowdsource information and technologies can be used for response-coordination when needed, and finally suggests related areas for future research.

Keywords Crowdsourcing · Event maps · Community maps · Crisis maps · Social media · Data mining · Machine learning · Humanitarian Aid and Disaster Relief (HADR)

G. Barbier
Arizona State University
E-mail: gbarbier@asu.edu

R. Zafarani
Arizona State University
E-mail: reza@asu.edu

H. Gao
Arizona State University
E-mail: huiji.gao@asu.edu

G. Fung
Arizona State University
E-mail: puicheonggabriel.fung@asu.edu

H. Liu
Arizona State University
E-mail: huanliu@asu.edu

1 Introduction

Crowdsourcing is an online production model that has emerged in recent years. The term crowdsourcing describes a new web-based business model that harnesses the creative solutions of a distributed network of individuals through what amounts to an open call for proposals [20]. According to Howe [20], crowdsourcing entails the work previously performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open calls by a company or institution. This can take the form of peer-production, but is also often undertaken by socially-networked individuals.

Crowdsourcing is often used to obtain solutions to a problem that are cheaper and superior in quality and quantity to those that are obtained from traditional professionals in the same industry. Successful examples of crowdsourcing include: Threadless[1] which has a crowd design a handful of original t-shirts each week usually selling out of stock, iStockphoto[2], where a crowd produces stock photography on par with professionals, and InnoCentive[3], where a crowd outperforms industry faster and cheaper than the top minds in the field. All three cases show that the solutions provided by crowdsourcing are better than the solutions provided by traditional problem-solving methods.

Crowdsourcing depends heavily on Web 2.0 technologies. Web 2.0 facilitates interactive information sharing, interoperability, and collaboration. In fact, Web 2.0 is the necessary technology that can help realize the wisdom of the crowd and coerce a mass of users into productive labor. With the advance of Web 2.0 and the advent of social media, even though different people are spread throughout different geographical locations and among a variety of cultural backgrounds, the web can facilitate the collection of information and distribution of problem solving in new ways. To this end, the web actually provides a medium for individuals around the world to gather in a single environment, regardless of their physical location [35].

The asynchronous nature of communication on the web makes possible the aggregate of disparate flows of ideas in one stream, continuously. Unlike synchronous communication where all participants must be present at the same time, asynchronous communication does not necessarily occur instantaneously. That is, individuals may participate at a time and location that is convenient for them [11].

The core concept in crowdsourcing has been around for some time. For example, open source software development can be considered a type of crowdsourcing. Popular software such as Firefox, Apache, and Linux are the results of crowdsourcing. User-generated content including Wikipedia, YouTube, Yahoo Answers, and social bookmarking are also good examples of productive crowdsourcing.

[1] http://www.threadless.com/
[2] http://www.istockphoto.com
[3] http://www2.innocentive.com/

Crowdsourced data can be used for a variety of situations to give people better insight into events that impact their communities. In some instances the uniting factor is a crisis; in other cases it could be an election or another public event. Crisis maps, also known as community maps [14], have emerged as a common mechanism for communicating information about such events. Crisis maps are one of the most interesting applications of crowdsourced data. These crisis maps can generate a massive amount of data, making it difficult to understand and prioritize all of the data that is generated from a crowd. For example, the devastating earthquake in Haiti during January 2010 generated over 13,500[4] crowdsourced messages (in nine categories). This data provides useful insights but the data can be even more valuable after pre-processing.

Crowdsourced applications can reveal important details during a crisis but have lacked the ability to efficiently provide a mechanism to help coordinate a response [14]. This is particularly a challenging issue when there are multiple responders with disparate goals, resources, and communication systems. Both government and non-government organizations (NGOs) provided resources for the response in Haiti. However, there was no common information system for coordination that could be shared by all of the groups providing resources for the response.

Government information systems often do not interface with external information systems due to security precautions, reliability concerns [33], or other unique interface requirements. Similarly, some NGOs may not want a government organization entangled in their information systems. However, new approaches to sharing information could create an environment that will produce synergistic results while avoiding traditional barriers to crisis response. A common coordination system could be shared between government agencies and NGOs. Using a common medium, government agencies from different countries and NGOs can leverage crowdsourced data and applications to efficiently manage logistic requests and resource distribution in the midst of a crisis.

This article highlights how applications for crowdsourcing are used during a crisis, investigates challenges and approaches associated with data mining crowdsourced information, discusses how crowdsourcing can be used to support data mining, details an approach inspired from crowdsourcing for Humanitarian Aid and Disaster Relief (HADR) coordination, and notes when crowdsourcing is not the best approach to a problem. Finally, challenges and suggested areas for future research are presented.

2 Background and Related Work

Sites such as Wikipedia[5], an online encyclopedia, and OpenStreetmap[6], a world map, provide large groups of people the opportunity and mechanisms

[4] http://www.noula.ht/
[5] http://www.wikipedia.org/
[6] http://www.openstreetmap.org/

to create products that could not be created by an individual alone. Jeff Howe coined the term crowdsourcing in his 2006 magazine article. Howe describes crowdsourcing, in part, as leveraging the "latent talent of the crowd" [20] to accomplish something. For crisis maps, the latent talent of the crowd is sensing and reporting what is happening around them. The effective summation of these reports, coupled with associated geospatial information, can give broad (resulting from the masses) and rapid (resulting from the medium) situational insights, not otherwise possible.

One of the most common applications used to implement a crisis map is the Ushahidi[7] platform. Sahana[8] offers another crisis management platform. The Ushahidi software is open source and freely available. Within minutes of installation, Ushahidi implementers can have a crisis map tailored for the circumstances including customized report categories and news feeds. When made available to the public, the crowd can submit reports, including images and videos, organized around a map of the area of interest that is publicly accessible via the internet. Crowds have multiple mechanisms for submitting reports including Short Message Service (SMS), smart phone applications, microblogs (such as Twitter[9]), and directly through the crisis map web site.

Independent of a crisis map, Microblogs are proving to be utile crowdsourcing tools [33,16]. In 2009 Twitter added an option to include geographic location information with users' microblog messages, known as Tweets. A large set of Tweets relevant to a particular event can be an excellent source of crowdsourced data. There are numerous proposals about ways to utilize data from Twitter, such as work collaboration [19], education and learning [15], collective wisdom and promotion [22], and emergent events identification [21].

Facebook is another popular application that can be used for crowdsourcing data like business and market analysis [26], urban planning [8], and product repository generation [9]. Facebook provides many applications such as designing surveys, opening forums for discussion, dropping a note to bring awareness to a topic, and creating interest groups. Facebook is a mechanism for building brands, calling people to action, or even introducing ideas.

Researchers and businesses are endeavoring to increase efficiencies and optimize gains from crowdsourced data. Scientists at the University of Colorado in Boulder have embarked on a project for Empowering the Public with Information in a Crisis (EPIC)[10]. One of the areas that *Project EPIC* is cultivating is Tweet standards. The goal is to make it easier for machines and people to process information. Project EPIC recommends hashtags[11] that can be habituated into tweets during a crisis. For example, a portion of the Tweet may include a hashtag followed by a description such as "#need food" to specify that food is needed by someone. This becomes valuable when coupled with ad-

[7] http://www.ushahidi.com/
[8] http://sahanafoundation.org/
[9] http://twitter.com/
[10] http://epic.cs.colorado.edu/
[11] http://twitter.pbworks.com/Hashtags - a label for Tweets prefixed with the # character.

ditional information that can be available in the tweet such as the username and location information. In the "#need food" example, "need" is tagged.

Some research has [39] has focussed on dealing with challenges surrounding crowdsource-provided labels for images. These challenges include handling cases when items are labeled with conflicting labels and labels produced by annotators with different skills. While tagging is a portion of the challenge that must be addressed with crisis map data, other elements must be addressed such as whether or not to report a message or discard it completely.

Crowdsourcing is a useful approach because it can be utilized by virtually any group for a variety of causes. The Defense Advanced Research Projects Agency (DARPA) [12] recently released ten helium balloons and challenged people to pinpoint their locations. The results attested that a collective group can solve time-sensitive problems. Additionally, crowdsourced data can be used to ascertain the general sentiment of a crowd. Applications such as *WeFeelFine*[12] use social media data to assess a variety of emotional states. Additionally, sentiment propagation has been investigated [41]. Similar approaches can be applied to crowdsourced data to better inform organizations about the emotional state of a crowd during a crisis. As an example, using the *Twitter Sentiment* website[13] with search terms "haiti earthquake" provides a set of statistics about the search term based on users' microblogs statements (Tweets).

Crowdsourcing certainly provides some attractive opportunities for gaining new insights into a variety of problems. However, in order to maximize the benefits of crowdsourced data care must be taken to prepare the data for processing and the right data mining approach needs to be considered for the specific problem at hand. These are especially important considerations to make when leveraging crowdsourced data for appropriate and efficient disaster response and coordination. In some some cases, leveraging crowdsourced data is not the right approach for solving the problems. Next we examine these points more closely and discuss areas we think are important for future work.

3 Preparing Data for Mining

Data mining methodologies and technologies can be applied to crowdsource data sets to better organize data generated from crowds and reveal additional information about events. Mining crowdsourced data can be performed by following a two-step procedure: *pre-processing* and *mining*. Due to the nature of crowdsourced data, a pre-processing step is necessary for the data to be ready for the mining step. After the pre-processing step, the data is "cleaner" and ready to be mined by various data mining and machine learning algorithms.

[12] http://www.wefeelfine.org/
[13] http://twittersentiment.appspot.com/

3.1 Preprocessing Data

Crowdsourced data in many cases comes in free-form text. However, most data mining and machine learning techniques require data to be in tabular format. In order to transform data to this format, various techniques from text mining [27] and Information Retrieval (IR) [4] are borrowed. We overview these techniques briefly.

- **Trust Assessment**: To validate the trustworthiness of crowdsourced data, various techniques could be employed. Voting techniques are proven to be successful in determining the trustworthiness of messages from various social media sites. For instance, on YouTube, users can provide feedback (thumbs up/thumbs down) for user comments. As a result, comments with too many negative feedbacks are automatically hidden. Ushahidi users can also vote on the credibility of an incident report. Another trust assessment technique is hierarchical trust management, where administrators at different levels manage various groups. These administrators moderate messages and information received from these users. Other techniques, such as reputation and trust modeling, can also be useful in this area [23].
- **Stop-word removal**: Stop-words are "words that occur frequently in text but have little meaning [27]." For example, the words "to, a, & the" are common stop words. These words can be safely removed prior to mining crowdsourced information because the mining algorithms will operate more efficiently without the stop-words, as the remaining words convey the most useful information.
- **Vectorizing**: As mentioned previously, tabular data is the preferred format for machine learning and data mining techniques. Converting text documents to vector space model tabular format is a well-known practice for vectorizing textual data. This technique, equipped with the power of measures such as the Term Frequency-Inverse Document Frequency (TF-IDF) [27], have proven to be useful for IR tasks.
- **Feature Selection/Feature Extraction** After the vectorizing step, the data is transformed to tabular format. Each entry in these vectors represents a feature [17]. These features can be filtered (feature selection) or combined (feature extraction) for generating a better mining performance [**?**,28].

Our experience in working with crowdsourced data (crisis map incident reports generated as a consequence of the 2010 Haiti earthquake) underscores the importance of pre-processing crowdsourced data. Our examination of the crowdsourced reports available from the Haiti earthquake[14] revealed that some reports might have been better placed in a different category than originally reported. This presents an accuracy problem for anyone who would like to use the crowdsourced data to respond to a crisis. For example, if requests are made for drinking water and submitted under under a natural hazard category (perhaps associated with a flooded area) the need for drinking water at a

[14] http://haiti.ushahidi.com/download/

Step	Data
Initial Data	`"1933","Help needed for orphanage at Delmas 31","2010-01-25 12:17:00","Delmas 31 (18.5614, -72.301)","I want to know how i can get some support for my Orphanage it is at Delmas 31 please tell what i have to do?Time:2010-01-25 12:17:04 Je veux savoir comment je peux obtenir de l?'aide pour mon orphelinat situ?? ?? Delmas 31? svp dites-moi ce que je dois faire ","3a. Penurie d'eau | Water shortage, 3d. Penurie d'aliments | Food Shortage, ","18.5614","-72.301","YES","NO"`
After Stop-word Removal	`Help orphanage "Delmas 31" support Water shortage Food`

Table 1 Pre-processing Crisis Map Incident Reports

particular location may not be noted in a reasonable amount of time. Preprocessing crowdsourced data can present available information better and serve as an important tool allowing governmental and non-governmental organizations to respond to a crisis in an organized and timely manner.

In our experiments with this data, all the previously discussed preproccesing steps were taken. Consider the incident report instance (textual data) received in Table 1.

As can be seen, the stop-word removal procedure cleans the data dramatically. The vectorizing and feature selection steps are not shown in this table. The vectorizing step takes the 7 remaining words and creates a large vector (size of this vector is the number of keywords in the corpus) where only 7 entries are one and the rest are zero. The feature section/extraction step takes these vectors as input and reduces the sparsity by removing the zero entries (see [6]).

3.2 Mining

After cleaning the data, machine learning or data mining techniques can be applied to the tabular data. Machine learning techniques can be divided into three parts: classification, clustering, and semi-supervised learning. In terms of classification techniques, support vector machine, bayesian approaches, and regression methods are the most celebrated methods [6]. For clustering, density based and spectral based clustering techniques are widely used. Finally, the Expectation Maximization (EM) method is an method of semi-supervised learning [6]. For data mining, sequential pattern mining techniques are most helpful.

Returning to our Haiti incident report, after pre-processing we are left with a set of vectors, whereby each vector represents an incident report. To categorize incidents so that similar incidents fall into the same categories, clustering techniques could be employed. k-means clustering has been applied to our dataset in order to categorize incidents. The value for k has been extracted using an X-means clustering algorithm that determines the number of clusters automatically [30].

4 Mining Data from Crowdsourcing

There are various data mining tasks that can be performed by means of crowdsourcing, namely: classification, clustering, semi-supervised learning, validation, and sampling. These tasks can become too complex to be done by employing automated techniques. In these situations, humans can still perform more accurately and efficiently than a machine. We overview the role that humans play for each task below and provide examples that demonstrate the human computational crowdsourcing power for data mining.

4.1 Classification

Classification can be performed within crowdsourced data. For instance, users can categorize documents or can assign labels, also known as classes (i.e., tags), to documents manually. This approach has been successfully tested in various domains. Examples include, but are not limited to: Yahoo! Directory, social bookmarking sites, and social news. In Digg, users assign categories and tags to submitted links, i.e., diggs. The wisdom of the crowd allows for more accurate categorization than any other machine learning algorithm. For instance, in Digg [3], more relevant tags are more likely to be assigned by a large percentage of users to an article. This reduces noise drastically. Another well-known example is when humans help solve a complicated problem where machines fail. CAPTCHAs and the technology behind them help solve the problem of digitizing handwritten text [38]. In CAPTCHAs, humans help classify text by labeling images of text that need to be digitized. Figure 1 depicts a CAPTCHA instance.

Fig. 1 CAPTCHA

4.2 Clustering

Clustering can also be performed using crowdsourced data. In general, many recent social networking sites seek aid from humans in order to create categories. For instance, on Balatarin, a Persian social news site, users automatically create categories. Users are allowed to choose one of these categories for their submitted material. This can be viewed as an instance of automatic clustering of submitted articles. Another example is the case of Twitter, where users form clusters (known as trends or trending topics) by assigning hash tags to their tweets. This facilitates fast retrieval when searching for tweets and again, adaptively clusters streaming tweets.

4.3 Semi-Supervised Learning

In semi-supervised learning, a learning algorithm is given a subset of labeled data and another subset consisting of unlabeled ones. It is then required to label the unlabeled set using the information acquired from the labeled set. Similar to previous tasks, semi-supervised learning can also be performed with crowdsourcing. Consider the example of Amazon Mechanical Turk[15]. As mentioned in their website,

> Amazon Mechanical Turk is a marketplace for work that requires human intelligence. The Mechanical Turk service gives businesses access to a diverse, on-demand, scalable workforce and gives workers a selection of thousands of tasks to complete whenever it's convenient.
>
> Amazon Mechanical Turk is based on the idea that there are still many things that human beings can do much more effectively than computers, such as identifying objects in a photo or video, performing data de-duplication, transcribing audio recordings, or researching data details. Traditionally, tasks like this have been accomplished by hiring a large temporary workforce (which is time consuming, expensive, and difficult to scale) or have gone undone.

In Mechanical Turks, humans are given examples of how a given task can be performed correctly and then asked to generalize and solve the same task for some other instance. This is a well-practiced labeling technique for complex data labeling tasks in the data mining field [32].

4.4 Validation

Similarly, humans can help validate the results of clustering, classification, or semi-supervised learning. The validation task is performed as follows:

1. The learning task (classification, clustering, or semi-supervised) is performed using an automated technique.

[15] https://www.mturk.com/mturk/welcome

2. The same task (probably on a smaller-scale) is performed by humans (crowdsourcing).
3. The results are compared with the automated method's outcome and the accuracy is calculated. This gives an estimate of the overall accuracy of the automated method on a larger scale.

In [3], authors propose a method for identifying influential bloggers in the blogosphere. They evaluate their automatic technique by comparing their results to the crowdsourced data generated on Digg. They assume that the number of diggs humans have assigned to the posts submitted by influential bloggers should be higher and based on this assumption validate their findings.

4.5 Sampling

One of the complex tasks in data mining and machine learning is sampling. Due to the scale of current datasets, it is required to obtain samples with sufficient information so that the hypotheses devised from the sample information can be easily generalized to larger datasets. However, the question that needs to be answered here is how the sampling distribution should be selected so that the information obtained from the sample is maximized. Humans have proven to be credible information samplers, and it turns out the criteria that they aim to maximize results in maximally informative samples [38].

5 Crowdsourced Data for Response Coordination

We are investigating how to leverage crowdsourced data to better enable Humanitarian Aid and Disaster Relief (HADR) responders to rapidly gain situational awareness and coordinate relief efforts from disparate sources. In [14] Goolsby presents the case for an "inter-agency" map to enable collaboration. We view this as the next logical step for leveraging crowdsourced information. Crisis-maps are a natural starting point for inter-agency collaboration. Agencies from different nations, with different means, and with a variety of resources can coordinate their response to a crisis through a shared, semi-open, platform [**?**]. Taking advantage of the latest information technology and available crowdsourced crisis data, the groups will be able to respond more effectively and efficiently from group to group. This *response group*, comprised of disparate responders, is certainly part of the crowd that is interested in the crisis event, but the response group differs from the rest of the crowd based on the simple fact that it has control over resources used for crisis response. For example, a response group might be comprised of military organizations from various countries, the International Red Cross[16], Doctors Without Borders[17], and components of the United Nations Office for the Coordination of Humanitarian Affairs (UN OCHA).

[16] http://www.icrc.org/
[17] http://www.doctorswithoutborders.org/

In [13], we define and discuss *groupsourcing*:

> "*Groupsourcing is intelligently using information provided by a sanctioned group comprised of individuals with disparate resources, goals, and capabilities. Essentially, the response group is performing crowdsourcing but specialized for the entire group and taking it a few steps further. Using information provided from the crowd and the sanctioned members of the group, a groupsourced response can be led, managed, and coordinated efficiently.*
>
> *In order to be more efficient and most effective in helping to resolve the crisis, the members of the response group must subscribe to the centralized administrative control of an information management system to ensure data integrity, data security, accuracy, and authentication will be realized for each member of the response group. One member of the response group can field the response coordination system ensuring equal access to other members of the response group.*"

It may be necessary to have representatives from two or more groups acting in an administrator role to facilitate trust and communications efficiency, especially when multiple languages and cultures are evident in the response group. In order to best address these requirements, the response coordination system should support at least five components: crowdsourcing, request collection, response, coordination, and statistics.

5.1 Crowdsourcing

In [29], Liu and Palen summarize 13 crisis-related mashups to derive some high-level design directions of next generation crisis support tools. One of the most famous crisis maps is "Ushahidi"[18]. It utilizes web 2.0 technologies to collect and visualize real-time public reports on a map via SMS, email, and the web. It was first developed to track reports of incidents of violence in Kenya[19] after the 2007 elections. After the 7.0 magnitude earthquake striking Haiti, Ushahidi launched "Haiti Live"[20] to gather the post-earthquake crisis response and recovery efforts in Haiti. Multitudes sent reports of missing people, destroyed buildings and requests for water, food, and medical equipment to "Haiti Live." Relief organizations and volunteer individuals responded to these requests based on the request urgency and their available resources. The crowdsourcing data provided government and non-government organizations with an abundant resource to gauge the situation.

Raw crowdsourcing data is problematic because of a large amount of noise in the data. In some cases people send multiple requests if one is not fulfilled. Reports may contain typing errors in keywords or be confusing due to the

[18] http://www.ushahidi.com/
[19] http://legacy.ushahidi.com/
[20] http://haiti.ushahidi.com/

stresses experienced in an emergency. The requester may even forget to specify the category of the request or type a wrong category. Malevolent attacks create fake requests to further impede the relief progress. Ushahidi adopted an open source platform named "SwiftRiver"[21] to help filter and verify real-time data from channels such as Twitter, SMS, Email and RSS feeds. Machine learning and data mining approaches can provide some assistance to clean up the noise amongst the original reports and re-classify requests into more accurate categories. A latent challenge with the raw data is the unstructured format of the reports. For example, some requesters may not specify the quantity in their report: "Please thirst is killing us in Delma 3 Idalina" ("Haiti Live" request #4008[22]), "The people what a real problem for food in Crois-des-Bouquets especially in the district of Santo 15 and 13." ("Haiti Live" request #3971[23]). Domain experts for a particular region and crisis will be required to help quantify the approximated quantity from this kind of requests.

Although crowdsourcing data can provide useful information to relief organizations, the crowdsourcing data and mechanisms we have analyzed to date are not yet sufficient for relief organizations to coordinate an efficient response. Certain crisis map implementations use a colors to indicate the report category and a different icon size to indicate the number of reports at a particular location. This visualization strategy is able to provide request distribution information to the relief organizations but is less helpful in relief effort. As the number of reports do not indicate any quantity information of requested resource, relief organizations still need to process the reports to understand the magnitude of the relief requirement. In order to reduce response time and facilitate the relief effort, a more efficient and effective visualization strategy is needed that provides not only the request category but also the magnitude of the request.

One approach for visualizing the magnitude of the requests is to extract the quantity and units of the request resources and visualize the quantified requests as nodes on a map. Nodes with different sizes indicate different request quantities at that location. Figure 2 demonstrates the visualization of raw crowdsourcing requests in the food category with labels on the nodes representing the number of reports at that location. Figure 3 demonstrates the visualization of the quantified requests of the same category, with labels on the nodes indicating the quantity of food requested at that location. For example, 170 indicates this request asks for 170 k-pounds of food in that region. Compared to the raw data visualization, the quantified resources map is more helpful.

With existing named-entity recognition [24,31] semantic similarity technologies [5,36], and language translation tools[24], important information can be auto-extracted from user generated reports. For example, consider again

[21] http://swift.ushahidi.com/
[22] http://haiti.ushahidi.com/reports/view/4008
[23] http://haiti.ushahidi.com/reports/view/3971
[24] http://code.google.com/apis/language/

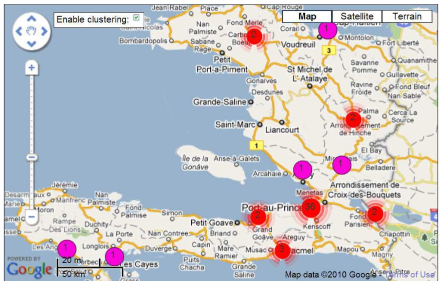

Fig. 2 Raw Crowdsourcing Requests

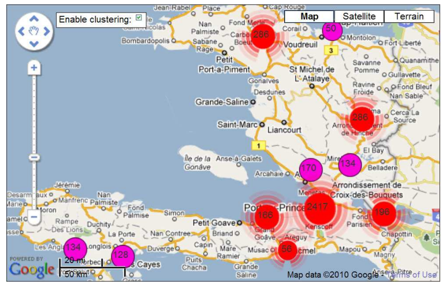

Fig. 3 Quantified Requests

<table>
<tr><th>Step</th><th>Data</th></tr>
<tr><td>Initial Data</td><td><code>"1933","Help needed for orphanage at Delmas 31","2010-01-25 12:17:00","Delmas 31 (18.5614, -72.301)","I want to know how i can get some support for my Orphanage it is at Delmas 31 please tell what i have to do?Time:2010-01-25 12:17:04 Je veux savoir comment je peux obtenir de l?'aide pour mon orphelinat situ?? ?? Delmas 31? svp dites-moi ce que je dois faire ","3a. Penurie d'eau | Water shortage, 3d. Penurie d'aliments | Food Shortage, ","18.5614","-72.301","YES","NO"</code></td></tr>
<tr><td>Named-entity recognition result</td><td><code>LOCATION: Delmas</code></td></tr>
<tr><td>Automatic language translation result</td><td><code>I want to know how I can get some? Help for my orphanage located? ? Delmas 31? please tell me what I should do</code></td></tr>
</table>

Table 2 Example Crisis Map Incident Report after automatic language translation

a message from the Haiti earthquake in table 2[25] [26] Applying named entity recognition automatically identifies the location as Delmas. Table 3 shows another example of extracting resource quantity information from a raw report. Using these techniques, relief organizations will be able to obtain, understand, and process requests more efficiently.

In addition to the crisis map reports, other social media data can provide situational awareness information about a region impacted by a crisis. For example, *Blogtrackers* [1] is an application that allows sociologists to analyze the blogosphere to study trends in blogs and even identify influential bloggers [2]. During the later stages of relief efforts, organizations can improve their understanding about a region using tools such as blogtrackers to better gauge public sentiment and leverage influential bloggers to assist with disseminating information important to the crowd.

[25] Results from http://translate.google.com using "Detect Language" feature

[26] Results from http://cogcomp.cs.illinois.edu/demo/ner/results.php online named entity recognition demonstration

<table>
<tr><th>Step</th><th>Data</th></tr>
<tr><td>Initial Data</td><td>"3933","IDP camp of 234 families has received no assistance, Cite Soleil","2010-03-29 00:05:00","Fountain Doulyn Number 6, cite Soleil","An IDP camp of 234 families in Abris Mochodem in Fountain in Cite Soleil has not yet received any assistance. There is an urgent need for tents as people are living under cloth sheets and without protection from the rain.","2b. Penurie d'eau | Water shortage, 2f. Sans courant | Power Outage, 2d. Refuge | Shelter needed, 2a. Penurie d'aliments | Food Shortage, 8a. IDP concentration, ","18.595033","-72.323917","YES","YES"</td></tr>
<tr><td>Named-entity recognition result</td><td><TIME>2010-03-29 00:05:00</TIME>
<LOCATION>Fountain Doulyn Number 6, cite Soleil</LOCATION>
<QUANTITY>234</QUANTITY>
<UNIT>FAMILY</UNIT>
<RESOURCE>TENT</RESOURCE>
<PRIORITY>URGENT</PRIORITY></td></tr>
</table>

Table 3 Example Crisis Map Incident Report before and after named-entity recognition

Crowdsourcing is the first step of disaster relief focusing on data collection. After preprocessing and cleaning up the noise in crowdsourced data, it can provide more valuable information to relief organizations than raw data. After responding on the scene, relief organizations will be most effective given the ability to communicate, collaborate, and work together. The response group should use crowdsourced data and tools used by the crisis population as one source to inform situational awareness. However, the group has additional requirements that must be met to ensure efficient relief operations and overcome known coordination barriers listed in [37]. The response group must be able to understand and prioritize relief needs, organize a response based on available resources, and track the overall progress of relief efforts.

5.2 Request collection

Request collection functionality enables the relief group to validate and quantify request for assistance using trusted data (from the relief group) while leveraging crowdsourced data for situational awareness. Crowdsourcing data cannot be used directly to manage request collection because crowdsourced data does not provide the granularity needed for planning, may not be accurate, and is not gauranteed to be updated after a relief action has been accomplished. To overcome these challenges, a host-agency model is a good choice. Using a host-agency model, requests are collected directly by a request collector who is working or serving in the disaster scene as a collecting agency. An administrator system can receive data and enable the relief-group to efficiently coordinate a response. Three kinds of data need to be collected in request collection:

- Request Quantity
  The collectors collect and record the approximated quantity of requested resources. Records will then be sent directly to the administrator system and can be visualized on map.
- Regional Situation Report
  The collector summarizes a periodic situation report based on the relief progress. Relief organizations can then make decisions based on these situation reports.
- Transportation infrastructure
  Transportation is a significant part of disaster relief. Local buildings and roads may be destroyed due to the disaster. It may be more difficult for the relief organizations to deliver relief supplies to a specified location in time without detailed information about the transportation infrastructure (e.g. road conditions, airfield status, and rail access). Reporting on transportation infrastructure enables relief organizations to verify their planned route before attempting to deliver relief supplies.

Request collectors are recruited and assigned to key locations by relief organizations.

5.3 Response

Currently, most crowdsourcing work focuses on how to collect data from crowds and present the latent, or summary, information about the event. A more advanced response model is required to improve relief efficiency. Request quantities can be extracted from crowdsourced data or generated directly through request collection. By mapping requests in different quantities to nodes with different sizes, we are able to generate a request pool and visualize requests on a map in the form of quantity, unit and resource type. By zooming in or zooming out on the map, nodes indicating the same type will merge or split based on their location closeness. We call nodes merged by other small nodes a “request cluster” and nodes that cannot split into smaller nodes a “singleton request.” This map is more intuitive for making decisions about relief efforts. Consider an organization has 1000 gallons of fresh water to contribute. The organization can zoom-in or zoom-out on the request map to seek a cluster with approximatly 1000 gallons of water in the request, while with crowdsourcing data only, the organization would have to assess every report to estimate the amount of requested resources before planning delivery. This new response model will help the relief organizations effectively distribute resources and minimize response time.

The request pool can also be utilized to form a request network. The node in this network still indicates a certain quantity of requested resources, while the edge between two nodes indicates the available routes from one to the other.

We can still visualize this network on a digital map, except that unavailable routes will be hidden on the map. The map is editable by users with certain access levels such as advanced request collectors. They can update the available routes on the map based on the damage to transportation infrastructure. Relief organizations will be able to respond to the requests by picking the most optimal transportation route.

5.4 Coordination for Improved Logistics

During an event, crisis maps can help relief organizations obtain situational awareness about the crisis area and quickly visualize crowdsourced data on a map. However, current crisis map applications do not provide relief organizations the tools needed to coordinate and cooperate with each other efficiently. Organizations working on their own may easily cause conflicts if more than one organization is trying to respond to the same requests simultaneously. In some cases organizations do not inform each other before taking action. Usually there exists an organizing entity to help coordinate relief efforts. However, organizations may have the perception that working with an organizing entity limits their autonomy and reduces their freedom to make decisions as desired. An inter-agency [14] approach will help address this problem. An inter-agency approach provides a functional layer between organizations, acting as an organizing entity. Each organization uses the inter-agency application to check the available requests (i.e. requests that haven't been fulfilled by other relief organizations) and responds to certain requests as desired. The inter-agency application only publishes requests that are still un-met and will not publish the information of other responders.

The inter-agency application can be used to help organize relief actions such as giving recommendations to organizations based on budget and coordination between two organizations as in a case when both organizations are willing to respond to one request. In addition to tracking whether or not opportunities are "available" or "unavailable", the inter-agency application can be used to track the status of a relief request such as "in process," "delivering," or "delivered" to mark the status of requests. By marking the status of requests, request statuses can be layered. Filled requests will layer down and disappear or transfer to the background gradually while open requests will layer up to highlight a growing need to relief organizations.

5.5 Statistics for a Clear Big Picture

With the inter-agency application, organizations may not be informed of other organizations' actions since relief organizations might only be concerned with the available requested resources. However, action information will be very useful to obtain a global view of relief efforts and help determine future steps in relief strategy. The statistical component collects such information and calculates statistics about the relief effort to help identify latent patterns. A statistical component provides information in at least three areas:

- Contributions from different organizations
  Helps organizations decide whether or not to recruit more relief organizations to participate in the disaster relief effort based on current needs and available resources.
- Spatiotemporal information about the requests
  By calculating statistics and analyzing the spatiotemporal patterns among requests, it is possible to analyze the disaster situation and also plan for future events.
- Distribution of resources by organization
  Different organizations may have different budgets on various resource types. From the distribution statistic, organizations can determine their impact on the relief effort.

5.6 Lead, Manage, and Respond

To enable the groupsourced response requires additional functionality be added to crisis maps applications. Specifically, *lead*, *manage*, and *respond* functions must be available. The five components of the groupsource application enable the response group to perform these essential functions.

Based on the crowdsourcing data that is available, the relief group has insights as to what people impacted by the crisis need and can then gauge population sentiment, assessing what needs to be emphasized and addressed during relief operations. Request information provides more detailed information to the relief group, enhancing the ability of group members to lead their individual efforts and providing a capability for the group to effectively lead the relief effort.

The response module provides the relief group with a consolidated ability to manage resources that are available for relief efforts and logistics, enabling the relief group to respond efficiently. Finally, the statistics module provides information about about how well the relief group is managing needs and leading relief efforts.

In an effort to implement a model of how a relief group-centered collaborative coordination system might be realized, we are developing the Arizona State University (ASU) Coordination Tracker (ACT) system. ACT will be an open system, leveraging crowdsourced data and providing groupsourced functionality. Unlike applications that are focussed on emergency operations, for example *WebEOC*[27], the primary focus of ACT is to facilitate Humanitarian Aid and Disaster Relief (HADR) efforts. ACT will implement the five essential components to enable a response group to lead, manage, and respond efficiently.

[27] http://www.esi911.com/esi/

5.7 Overcoming Coordination Barriers

Any system designed to support groupsourcing should strive to address the coordination barriers identified in [37] including:

- "*The perception that coordination will limit autonomy and that the freedom to make decisions and run programmes as desired will be circumscribed.*" Although members of the response group need to subscribe allowing one or more members to manage the groupsource application, an open request collection process and response module allows an individual group to retain autonomy by enabling the group to choose how and when they will participate in the relief effort.
- "*Too many decision-makers or too many organizations involved which will complicate the process and make consensus, or at least agreement, too difficult to achieve.*" We believe by sharing information amongst the group and providing choices for specific relief actions, individual groups will recognize the benefit of open collaboration.
- "*Different expectations or beliefs about what is important, a priority, or the right thing to do in a given situation.*" We do not think our approach will eliminate this barrier completely. However, we believe that by providing a common system where relief organizations can view the big picture, have choice in response options, and collaborate openly; relief organizations will be better equipped to overcome this barrier for any particular crisis.
- "*Lack of resources to devote to coordination or coordination seen as a low priority given limited time and resources.*" Since the groupsource system will rely on commercially available hardware (most likely already owned by relief organizations) and openly accessible software (via the internet or consumer based communications mechanisms such as cell phones), relief organizations will find that through collaborating with other organizations using this framework will result in more efficient efforts than previously possible.
- "*Limited field-based decision-making authority such that no decisions can be made without HQ approval thus resulting in delays or having an agreement overturned.*" Our approach makes it easier for field representatives to communicate and share up-to-date information with headquarters to better address this challenge.
- "*Staff turnover where new staff lacks a commitment to coordination or are unaware of coordination agreements.*" We believe that our approach of using commercially available systems will allow incoming staff to become proficient quickly by having access to information before arriving in the field. Additionally, the system will track resources and relief organization status along with the other members of the relief group, allowing new staff personnel to transition as quickly as possible.
- "*Unilateral actions that ignore established coordination mechanisms of the coordination body whether by donors or member organizations.*" By subscribing to a centralized process with synergistic opportunities for collaboration, we believe unilateral actions will be minimized because uncoordi-

nated unilateral actions will be easily identified and highlighted through both crowdsourcing and groupsourcing. Additionally, we believe that this new approach to collaboration will reduce tendencies for unilateral actions because the relief group will be able to operate more efficiently, have better situational awareness, and better insight overall.

- "*Ineffectual or inappropriate coordination leadership, for example, when the coordination body exercises autocratic leadership and imposes decisions on others without a transparent process of involvement.*" We believe the members of the response group must subscribe to the centralized administrative control of an information management system to ensure data integrity, data security, accuracy, and authentication to be realized for each member of the response group. We also believe that our approach provides the most transparent process for relief organizations.
- "*A coordination process that is not working well, has unclear objectives, and is seen to waste time without obvious benefits to those participating in it.*" Our approach will allow relief organizations to track needs, responses, and progress. We believe this will highlight the obvious benefits of collaboration when each organization can see the benefit of their contribution to the relief effort.

6 When Not To

We can ascertain the benefit of crowdsourcing in a number of situations, such as crisis response, work collaboration, and collective wisdom for decision making. However, crowdsourcing may not be suitable for tackling all tasks in all situations. Simply adopting the idea of crowdsourcing without cogitating whether it is suitable to the problem's domain or to the desired results is risky. Overusing crowdsourcing will only lead to solving a task less effectively, and efficiently, or may even obtain unexpectedly poor results.

Some articles [10,18,25,34] present critical reviews about crowdsourcing and suggest situations where and when crowdsourcing should not be employed. In summary, crowsourcing should be obviated when:

1. The knowledge needed is specific. One of the most appealing properties of crowdsourcing is collective wisdom – everyone contributes a portion about what is known and the crowd as a whole will likely yield better information than if an individual had provided it alone. Unfortunately, if the problem that needs to be addressed or the knowledge sought is too specific and not known by most of the general audiences, then crowdsourcing may not be the proper approach. Time will be wasted on explaining the problem domain, identifying the target crowd, and filtering the ineffectual information.
2. A problem is not well-defined. In most of the situations, crowdsourcing is used to solve problems that have a single objective and a clear focus, such as the DARPA balloon challenge [12], Wikipedia entries, etc. If the problem has multiple objectives or the problem is not well-defined, then

crowdsourcing should not be used, unless one can divide the problem into a series of sub-problems.
3. The problem itself cannot attract people's attention. Experience with Wikipedia indicates that people will contribute to things that are interesting to them. In fact, only a small proportion of people create or share content, a few active creators or editors, with the majority of people not participating at all [25]. As a result, if the problem at hand is not interesting to most people, then crowdsourcing may not yield the answer, as it would be difficult to collect enough information. [7] provides an excellent discussion on how crowdsourcing works on an everyday basis.
4. The problem is not fully disclosed. Using the public as a crowdsource is obviously not going to protect a problem that is not intended for open dissemination because it would be impossible to protect confidential information. Crowdsourcing something that is meant to be a confidential compromises the information.
5. The problem is a large project that requires dedicated and long term resources. For example, in most large and prominent open source software development projects, project leaders and members are paid by their employers to lead their projects because the software is beneficial to their employers, even though the software is not owned by the employer [40]. With such a reward model, long term commitment becomes easier to achieve and well-written documentations will be made available. Even when the leader or some members of the project leave the team, the project continues.

Additionally, crowdsourcing is usually performed by workers not associated with a company or specific organization. Since there is no contract involved in crowdsourcing, the quality of work cannot be guaranteed. As such, when we have to solve a problem through crowdsourcing, the crowd should be vast, so that the chance for obtaining a high quality result is increased. For example, in the situation of Ushahidi, if there are only a small number of people willing to utilize Ushahidi to report an incident, there may be many false alarms and insufficient information to assess a crisis or event. On the other hand, if most people are willing to report incidents to Ushahidi during a crisis, a much higher quality result is obtained. For example, if a location received multiple reports, it is more likely that the location does have some problems and requires a response.

It is worth noting that crowdsourcing can have a negative impact on an industry. For instance, three of the successful crowdsourcing applications (Threadless, iStockPhoto and InnoCentive) take revenue from traditioanl professional elite in that industry. Threadless designers earn far less than professional clothing designers would if the design work were outsourced to them. iStockphoto photographers earn only $5 to $10 for each photo each time someone downloads their photos, but professional stock photographers can earn hundreds or thousands of dollars for each copy of their photos. InnoCentive solvers may win very large awards, but when compared with the cost of an in-house researcher, earn much less.

While crowdsourcing can help us solve a number of problems more effectively, it may not be appropriate to use in all situations. Before deciding whether we should use crowdsourcing, we should understand what we want to achieve and consider whether or not crowdsourcing will yield sufficient advantages for a particular situation.

7 Conclusions and Future Work

Crowdsourced data has emerged as a valuable source of information about crisis events. We believe crowdsourced information can be even more useful after data processing techniques are applied to clean up and organize the data for end users. However, additional work is needed to validate our approach. A variety of data mining techniques can be applied to crowsdsourced data including trust assessment, stop-word removal, vectorizing, and feature selection.

Relief organizations are a subset of the crowd interested in or affected by a particular crisis. Relief organizations acting collaboratively as a relief group can supplement crowdsourced data with groupsource information. To maximize efficiency and situational awareness for the the relief group, an integrated response coordination system should support crowdsourcing, request collection, response, coordination, and statistics. As an integrated system is realized, future efforts should investigate how well a groupsourced crisis coordination system will overcome traditional barriers to collaboration during relief efforts.

Many challenging research questions remain to be addressed. The crowdsourced data is often noisy and filled with repetitive entries without much relevance. Denoising is an imperative process to filter out and clean up data for analysis. In order to facilitate collaboration between organizations, we could provide a recommendation module to allow the involved organizations to perform scenario planning based on their private capabilities before committing to physical contribution. Their private capabilities include budget and priority. Other factors include the emergency level of requests, delivery cost, and the most optimized route to deliver, and so on. One of the key complaints with donors is that they are not informed of fulfilment: what is delivered, how much of the need is met, who gets what. The monitoring module can address this issue and help smooth and improve coordinated disaster relief. We can also leverage the spatio-temporal information of previous requests and their corresponding response histories to anticipate potential needs in nearby areas for speedy HADR response.

Acknowledgements The authors wish to acknowledge the members of the Arizona State University, Data Mining and Machine Learning laboratory for their motivating influence and thought-inspiring comments and questions with reference to this topic. This work, in particular, the content of Section 5, was inspired by and based on an ongoing project "ASU Coordination Tracker (ACT) for Disaster Relief". This work was funded, in part, by the Office of Naval Research (ONR), the Air Force Office of Scientific Research (AFOSR), and the OSD-T&E (Office of Secretary Defense-Test and Evaluation), DefenseWide/PE0601120D8Z

National Defense Education Program (NDEP)/BA-1, Basic Research; SMART Program Office, www.asee.org/fellowships/smart, Grant Number N00244-09-1-0081.

References

1. Agarwal, N., Kumar, S., Liu, H., Woodward, M.: Blogtrackers: A tool for sociologists to track and analyze blogosphere. In: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media (ICWSM) (2009). URL http://www.public.asu.edu/~huanliu/papers/icwsm09k.pdf
2. Agarwal, N., Liu, H.: Modeling and Data Mining in Blogosphere, *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 1. Morgan and Claypool (2009). URL http://www.morganclaypool.com/doi/abs/10.2200/S00213ED1V01Y200907DMK001
3. Agarwal, N., Liu, H., Tang, L., Yu, P.: Identifying the influential bloggers in a community. In: Proceedings of the international conference on Web search and web data mining, pp. 207–218. ACM (2008)
4. Baeza-Yates, R., Ribeiro-Neto, B., et al.: Modern information retrieval, vol. 463. ACM press New York (1999)
5. Banerjee, S., Pedersen, T.: An adapted lesk algorithm for word sense disambiguation using wordnet. Computational Linguistics and Intelligent Text Processing pp. 117–171 (2002)
6. Bishop, C., service), S.O.: Pattern recognition and machine learning, vol. 4. Springer New York (2006)
7. Blodget, H.: Who the hell writes wikipedia, anyway? (2009). URL http://www.businessinsider.com/2009/1/who-the-hell-writes-wikipedia-anyway
8. Brabham, D., Sanchez, T., Bartholomew, K.: Crowdsourcing Public Participation in Transit Planning: Preliminary Results from the Next Stop Design Case. Transportation Research Board (2009)
9. Budde, A., Michahelles, F.: Towards an open product repository using playful crowdsourcing. Digitale Soziale Netze@ Jahrestagung GI, Leipzig (2010)
10. Campbell, M.: Innovation: The sinister powers of crowdsourcing. http://www.newscientist.com/article/dn18315-innovation-the-sinister-powers-of-crowdsourcing.html (2009)
11. Cristian, F.: Synchronous and asynchronous. Communications of the ACM **39**(4), 88–97 (1996)
12. DARPA: Darpa network challenge
13. Gao, H., Barbier, G., Goolsby, R.: Harnessing the crowdsourcing power of social media for disaster relief. Intelligent Systems, IEEE **26**(3), 10 –14 (2011). DOI 10.1109/MIS.2011.52
14. Goolsby, R.: Social media as crisis platform: The future of community maps/crisis maps. ACM Trans. Intell. Syst. Technol. **1**(1), 1–11 (2010). DOI http://doi.acm.org/10.1145/1858948.1858955
15. Grosseck, G., Holotescu, C.: Can we use Twitter for educational activities. In: 4th International Scientific Conference, eLearning and Software for Education, Bucharest, Romania (2008)
16. Grossman, L.: Iran protests: Twitter, the medium of the movement. Time (June 17, 2009). URL http://www.time.com/time/world/article/0,8599,1905125,00.html
17. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann (2006)
18. Harkin, J.: The trouble with twitter. http://www.guardian.co.uk/commentisfree/2009/dec/29/trouble-twitter-social-networking-banality (2009)
19. Honey, C., Herring, S.: Beyond microblogging: Conversation and collaboration via Twitter. In: System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on, pp. 1–10. IEEE (2009)
20. Howe, J.: The rise of crowdsourcing. Wired **14.06**, Retrieved 2010–10–04 (2006)
21. Hughes, A., Palen, L.: Twitter adoption and use in mass convergence and emergency events. International Journal of Emergency Management **6**(3), 248–260 (2009)

22. Jansen, B., Zhang, M., Sobel, K., Chowdury, A.: Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology **60**(11), 2169–2188 (2009)
23. Jusang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decision Support Systems **43**(2), 618–644 (2007)
24. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: International Conference on Machine Learning, pp. 282–289. Citeseer (2001)
25. Li, C.: Forrester's new social technographics report (2007). URL `http://forrester.typepad.com/groundswell/2007/04/forresters_new_.html`
26. Libert, B., Spector, J.: We are smarter than me: How to unleash the power of crowds in your business. Wharton School Publishing (2007)
27. Liu, B.: Web Data Mining. Springer (2007)
28. Liu, H., Yu, L.: Toward integrating feature selection algorithms for classification and clustering. IEEE Trans. on Knowledge and Data Engineering **17**(3), 1–12 (2005)
29. Liu, S., Palen, L.: Spatiotemporal Mashups: A Survey of Current Tools to Inform Next Generation Crisis Support. In: Proceedings of the 6th International ISCRAM Conference, Gothenburg, Sweden (2009)
30. Pelleg., D., Moore, A.: X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In: Proceedings of the Seventeenth International Conference on Machine Learning, pp. 727–734 (2008)
31. Rabiner, L.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)
32. Sorokin, A., Forsyth, D.: Utility data annotation with amazon mechanical turk. In: Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on, pp. 1–8. IEEE (2008)
33. Spellman, J.: Heading off disaster, one tweet at a time (2010). URL `http://www.cnn.com/2010/TECH/social.media/09/22/natural.disasters.social.media/index.html?hpt=Sbin`. Turner Broadcasting System, Inc.
34. Surowiecki, J.: G20 summit: How the bandwagon wrecked the wisdom of market crowds. http://www.guardian.co.uk/commentisfree/2009/mar/31/james-surowiecki-comment-global-economy (2009)
35. Terranova, T.: Network culture: Politics for the information Age. Pluto Press, London (2004)
36. Turney, P.: Mining the web for synonyms: Pmi-ir versus lsa on toefl. In: Proceedings of the twelfth european conference on machine learning (ecml-2001), pp. 491–502 (2001)
37. UNOCHA: United Nations disaster assessment and coordination (UNDAC) handbook. Electronic (2006). URL `http://ochaonline.un.org/OCHAHome/AboutUs/Coordination/UNDACSystem/UNDACHandbook/tabid/6012/language/en-US/Default.aspx`
38. Von Ahn, L.: Human computation. In: Proceedings of the 4th international conference on Knowledge capture, pp. 5–6. ACM (2007)
39. Welinder, P., Perona, P.: Online crowdsourcing: Rating annotators and obtaining cost-effective labels. pp. 25 –32 (2010). DOI 10.1109/CVPRW.2010.5543189
40. Woods, D.: The commercial bear hug of open source (2008). URL `http://www.forbes.com/2008/08/17/cio-open-source-tech-cio-cx\_dw\_0818open.html`
41. Zafarani, R., Cole, W.D., Liu, H.: Sentiment propagation in social networks: A case study in live journal. In: S.K. Chai, J.J. Salerno, P.L. Mabry (eds.) Advances in Social Computing, Lecture Notes in Computer Science, pp. 413–420. Springer (2010)