Groupsourcing: Distributed Problem Solving Using Social Networks

Jon Chamberlain
School of Computer Science and Electronic Engineering
University of Essex
Wivenhoe Park, Colchester CO4 3SQ, England
jchamb@essex.ac.uk

Abstract

Crowdsourcing and citizen science have established themselves in the mainstream of research methodology in recent years, employing a variety of methods to solve problems using human computation. An approach described here, termed "groupsourcing", uses social networks to present problems and collect solutions. This paper details a method for archiving social network messages and investigates messages containing an image classification task in the domain of marine biology. In comparison to other methods, groupsourcing offers a high accuracy, data-driven and low cost approach.

Introduction

Crowdsourcing has established itself in the mainstream of research methodology in recent years, using a variety of methods to engage humans to solve problems that computers, as yet, cannot solve. Whilst the concept of **human computation** (von Ahn 2006) goes some way towards solving problems, it also introduces new challenges for researchers, not least how to deal with human psychology. Issues of participant recruitment and incentivisation are significant and many projects do not live up to expectations because human effort cannot be acquired in the same way as machines. Three variations of collaboration over the Internet that have been successful can be distinguished by the motivations of the participants.

In the first approach the motivation for the users to participate already exists. This could be because the user is **inherently interested** in contributing, such as Wikipedia (Nov 2007), or because users need to accomplish a different task, for example the reCAPTCHA authentication system (von Ahn et al. 2008).

Many human computation tasks are neither interesting nor easy to integrate into another system, so a second approach to crowdsourcing called **microworking** was developed, for example Amazon Mechanical Turk (Kittur, Chi, and Suh 2008). Participants are paid small amounts of money to complete HITs (Human Intelligence Tasks). Simple tasks can be completed very quickly (Snow et al. 2008), however this approach cannot be scaled up for large data collection efforts due to the cost. Issues of ethics and workers' rights have also been raised (Fort, Adda, and Cohen 2011).

A third approach is to entertain the user whilst they complete tasks, typically using games or gamification. The purposeful games or **games-with-a-purpose (GWAP)** approach has been used for many different types of crowdsourced data collection including text, image, video and audio annotation, biomedical applications, transcription, search and social bookmarking (Chamberlain et al. 2013).

Projects that do not have the budget to recruit users on a large scale are reliant on accessing existing user groups. Social networks such as Facebook, LinkedIn and Flickr offer access to large user communities through integrated software applications and/or a back-end API. As social networks mature the software is utilised in different ways, with decentralised and unevenly distributed organisation of content, similar to how Wikipedia users create pages of dictionary content. This can be seen as an example of the broad term **collective intelligence** (Malone, Laubacher, and Dellarocas 2009).

Increasingly social networks are being used to organise data, to pose problems, and to connect with people who may have solutions that can be contributed in a simple and socially convenient fashion. **Citizen science**, where members of the public contribute knowledge to scientific endeavours, is an established research methodology and Facebook has been used as a way of connecting professional scientists and amateur enthusiasts with considerable success (Sidlauskas et al. 2011). However there are drawbacks with this method of knowledge sharing and problem solving: data may be lost to people interested in it in the future and it is often not accessible in a simple way, for example, with a search engine. This paper investigates this distributed problem solving on social networks and presents:

- a definition for the term "groupsourcing";
- a method for archiving social network messages;
- an analysis of task distribution amongst groups;
- an analysis of user demographics and workload;
- accuracy compared with other crowdsourcing methods.

Groupsourcing: A Definition

The utility of human computation became apparent when it was proposed to take a job traditionally performed by a designated employee and outsource it to an undefined large group of Internet users through an open call. This approach, called **crowdsourcing** (Howe 2008), revolutionised the way traditional tasks could be completed and made new tasks possible that were previously inconceivable due to cost or labour limitations.

In a similar way tasks can be completed by groups of users of social networking websites that are self-organised and decentralised. The tasks are created by the users, so they are intrinsically motivated to participate. The social nature of the groups allow users to connect with others of similar interests, with the reward being able to have their problem solved or to benefit from the problem being solved. Social media is entertaining and the natural language of the interface allows users to express their emotions, appreciation, frustration, etc. The combination of these motivations that relate directly to motivations of crowdsourcing generally (Malone, Laubacher, and Dellarocas 2009; Chamberlain et al. 2013) may explain why this approach has evolved from the users themselves.

Thus, a definition for **groupsourcing** is proposed as *completing a task using a group of intrinsically motivated people of varying expertise connected through a social network*.

This is more general definition than has been proposed before in relation to crowdsourcing disaster relief efforts (Gao et al. 2011) and could be applied to other question answering (QA) and opinion collection systems such as YahooAnswers[1], StackOverflow[2] and OpinionSpace (Faridani et al. 2010). It combines 3 central principles of crowdsourcing (crowd wisdom, creation and voting) (Howe 2008) and incorporates concepts of groupworking and group dynamics found in social psychology research (Forsyth 2005). The approach is also similar to crowd-powered websites such as iStockphoto[3] or Threadless[4], where the creation and validation of content and metadata is managed by the users.

Groups

A **group** in this context is a feature of a social network that allows a small subset of users to communicate through a shared message system. Groups are initially set up in response to the needs of a few people and the community grows as news from the group is proliferated around the network in feeds and user activity. The group title, description and "pinned" posts usually give clear indications as to whom the group is aimed at and for what purpose. This paper focuses on three types of group motivation that were considered likely to contain examples of problem solving:

1. Task Request (TR) - groups where users are encouraged to post messages with a task, e.g., *ID Please (Marine Creature Identification)*
2. Media Gallery (MG) - groups where users are encouraged to share media (image and video) for its artistic merit, e.g., *Underwater Macro Photographers*
3. Knowledge Sharing (KS) - groups used for coordination of activities or for distributing knowledge, research and news, e.g., *British Marine Life Study Society*

Category	Content
QUESTION	What is this?
CHECK	Is this Chromodoris magnifica?
NEUTRAL	Great photo from the trip!
ASSERTION	This is Chromodoris magnifica

Table 1: Categories of posts with examples of content, conditional on inquisition (question or statement) and data load (in this case the scientific name of a species in the image).

Category	Message	Reply
None	NEUTRAL	NEUTRAL
Unresolved	NEUTRAL	QUESTION
	QUESTION	QUESTION or NEUTRAL
Implied	NEUTRAL	CHECK or ASSERTION
	ASSERTION	Any
Suggestion	CHECK	Any
Resolved	QUESTION	CHECK or ASSERTION

Table 2: Categories of threads when viewed as a task with solutions.

Groups can also be categorised into those that are specific to a topic or subject (-S) and those that are non-specific or generalist (-G).

Threads, messages and replies

The thread of a typical post on a social network such as Facebook (see Figure 1) is structured:

1. A user posts a message.
2. Users (including the first user) can post a reply.
3. Users can "like" the message and/or replies including their own posts.

Messages posted to a group on Facebook can be one of 6 types: photo; link (URL); video; a question (in the form of an online poll); a scheduled event; or just simply text (status)[5] although the majority of messages are either "photo", "link" or "status" (see Figure 2).

For the purposes of this investigation messages and replies are categorised by **inquisition** (question or statement) and **data load** (a solution to the task – see Table 1), although more detailed schemas (Bunt et al. 2012) and richer feature sets (Agichtein et al. 2008) have been used to describe QA dialogue. The message and its replies form a thread that relates to what has been posted (photo, link, etc.). The thread may contain solutions (or related data) to tasks, irrespective of whether the poster posed a question in the original message, as other users might augment or correct the posts (see Table 2).

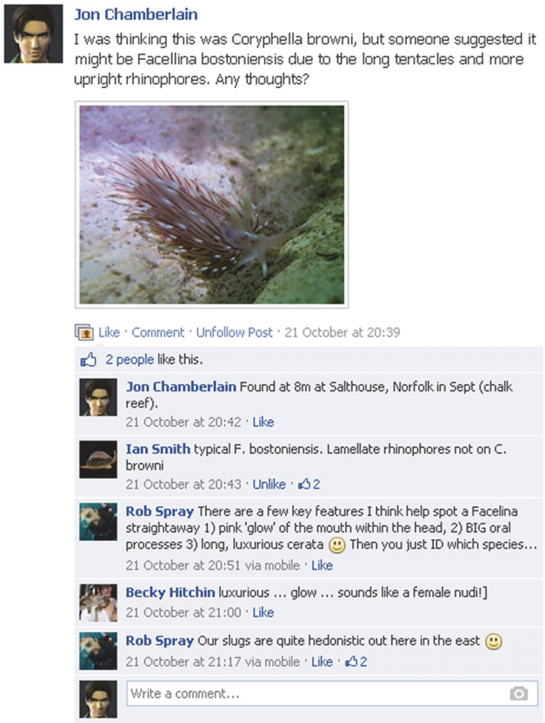
[1] https://uk.answers.yahoo.com

[2] http://www.stackoverflow.com

[3] http://www.istockphoto.com

[4] https://www.threadless.com

[5] http://fbrep.com//SMB/Page_Post_Best_Practices.pdf



Jon Chamberlain
I was thinking this was Coryphella browni, but someone suggested it might be Facellina bostoniensis due to the long tentacles and more upright rhinophores. Any thoughts?

Jon Chamberlain Found at 8m at Salthouse, Norfolk in Sept (chalk reef).
21 October at 20:42 · Like

Ian Smith typical F. bostoniensis. Lamellate rhinophores not on C. browni
21 October at 20:43 · Unlike · 👍2

Rob Spray There are a few key features I think help spot a Facelina straightaway 1) pink 'glow' of the mouth within the head, 2) BIG oral processes 3) long, luxurious cerata 🙂 Then you just ID which species...
21 October at 20:51 via mobile · Like

Becky Hitchin luxurious ... glow ... sounds like a female nudi!]
21 October at 21:00 · Like

Rob Spray Our slugs are quite hedonistic out here in the east 🙂
21 October at 21:17 via mobile · Like · 👍2

Figure 1: Detail of a typical message containing an image classification task posted on Facebook.

Message Archiving

In order to analyse the problem solving capabilities of social networks a pipeline to cache messages from Facebook groups was written in PHP and JavaScript and deployed on a live server. The software makes a request for a group's messages via the Facebook Graph API.[6] The call specifies the maximum number of messages to return (in date order, newest first) and the API returns a JSON encoded list of messages and metadata, termed here a **corpus**. The corpus is stored in JSON format in a MySQL database along with data about the group, such as the owner, title, description and privacy settings.

Each corpus contains a pagination link that is used to call sets of messages from a group. Pagination is used to minimise server load in processing large groups (avoiding timeout issues) and to circumvent Facebook's maximum message per call limit (500 messages). The software iterates through a group's messages from the latest message to the first message ever posted. The process of storing corpora from a group is termed here a **capture**.

The Facebook API was also used to find the gender of the each user, although users do not have to declare a gender or be truthful in their declaration. This was transformed into an anonymous database so users cannot directly be associated with the data held on the database. This use of data is in line with Facebook's Data Use Policy.[7]

Image Classification Task

The task investigated here is image classification (where objects in an image are identified) in the domain of marine biology. In this case the annotations are open (can be any text), although are later normalised to an ontology, and apply to the whole image. Region annotation, where parts of an image are annotated, is more complex and not fully supported by social networks yet. This task is particularly suited for social media because many sites are built around an image-text layout.

Automatic image annotation is an active area of research (Lu and Weng 2007) with specific industry supported tracks, such as Yahoo's Flickr-tag challenge at ACM Multimedia 2013.[8] Image annotation using crowds has also been an area of research (von Ahn and Dabbish 2004), however user participation has always been a limitation.

The power of mobilising a crowd to examine images on a large scale was pioneered by the search for sailor and computer scientist Jim Gray in 2007[9] and most recently seen with the disappearance of Malaysia Airlines flight MH370 in 2014.[10] Millions of users analysed satellite imagery, tagging anything that looked like wreckage, life rafts and oil slicks, with interesting images being passed on to experts.

Some citizen science projects get members of the public to classify objects in images taken from ROVs (Remotely Operated Vehicles)[11] [12] [13], whilst others require the users to supply the source data as well as the classification.[14] [15] [16] The latter has been less active due to technical constraints (the users need to be trained in SCUBA diving and have underwater photographic equipment) but empowered users to have their images identified by experts and contribute to scientific endeavours. The quality of citizen scientist generated data has been shown to be comparable to that generated by experts when producing taxonomic lists (Holt et al. 2013) even when the task is not trivial (He, van Ossenbruggen, and de Vries 2013).

Image classification in a QA format is common in marine biology and SCUBA diving forums[17] but suffers from not having a broad enough community of users to answer the questions. The image classification task in groupsourcing follows a similar QA dialogue style where threads may contain true tasks (where a question is asked and is answered) or implied tasks (where the post is augmented with additional data) and can be viewed as examples of distributed human computation problem solving.

[6] https://developers.facebook.com/docs/graph-api

[7] https://www.facebook.com/full_data_use_policy (15/11/2013)

[8] http://acmmm13.org/submissions/call-for-multimedia-grand-challenge-solutions/yahoo-large-scale-flickr-tag-image-classification-challenge

[9] http://www.wired.com/techbiz/people/magazine/15-08/ff_jimgray

[10] http://www.tomnod.com/nod/challenge/mh370_indian_ocean

[11] http://www.planktonportal.org

[12] http://www.seafloorexplorer.org

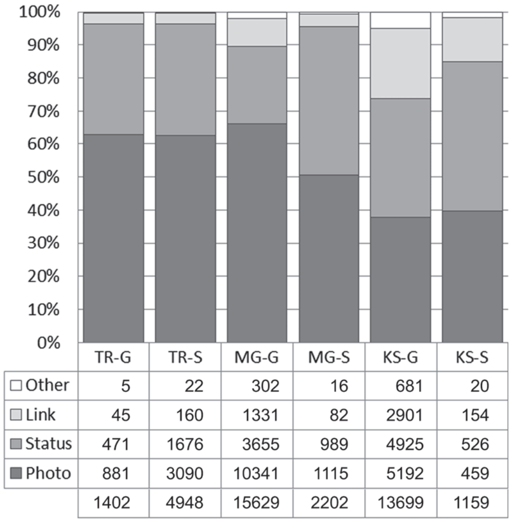
[13] http://www.subseaobservers.com

[14] http://www.projectnoah.org

[15] http://www.arkive.org

[16] http://www.brc.ac.uk/irecord

[17] http://www.scubaboard.com/forums/name-critter



	TR-G	TR-S	MG-G	MG-S	KS-G	KS-S
□ Other	5	22	302	16	681	20
□ Link	45	160	1331	82	2901	154
▨ Status	471	1676	3655	989	4925	526
■ Photo	881	3090	10341	1115	5192	459
	1402	4948	15629	2202	13699	1159

Figure 2: Distribution of thread types by group category.

Data Analysis

Facebook has a vast resource of uploaded images from its community of users, with over 250 billion images, and a further 350 million posted every day. Images of things (rather than people or places) that have been given captions by users only represents 1% of this data, but it is still of the order of 2.6 billion images.[18]

A proportion of these images are posted to groups where the classification task might be found. These groups were identified using the inbuilt search functionality, group recommendations and checking the group membership of prominent users in groups already found. Only groups that were sufficiently mature were selected (over 50 messages and 50 members) and were categorised according to purpose and generality (see *Groups*).[19] The total cached message database includes 34 groups containing 39,039 threads.

Analysis of a random sample of 1,000 messages from the corpus showed a rapid drop in replies to messages after 4 weeks. Therefore, for the purposes of analysing thread activity, all messages less than 8 weeks old from the date of capture were ignored to reduce any bias in message activity of newly posted and currently active messages.

Distribution of message type

The Task Request (TR) and Media Gallery (MG) groups have more photo type messages posted in them compared to Knowledge Sharing (KS) groups both in the general and topic-specific categories ($p<0.05$, z-test). This is not surprising as the primary motivation for posting a message in TR and MG groups (seeking an identification or showing off a picture, respectively) requires an image to be attached. The KS groups show a more even spread of message types as motivations for posting (arranging meetings, sharing research, posting information, etc.) do not require an image. This makes TR and MG groups better places to look for image classification tasks.

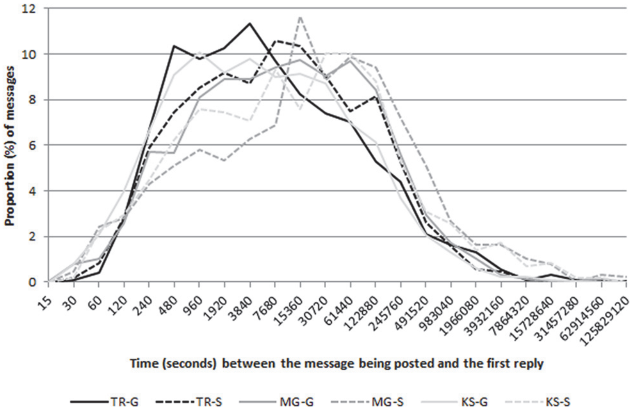


Figure 3: Response time (seconds, log scaled) for a message.

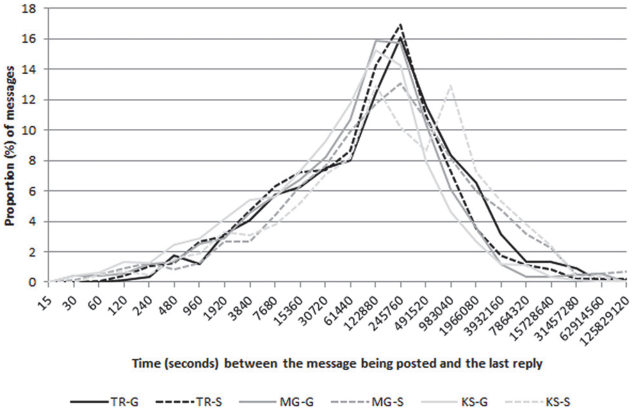


Figure 4: Lifespan (seconds, log scaled) of a message thread.

Thread response time, lifespan and activity

The time to the first response (response time) and time to the last response (lifespan) were plotted on frequency graphs (see Table 3 and Figures 3 and 4). 5-10% of messages receive a reply in 8 minutes. The proportion of messages with replies beyond 1092:16:00 (6.5 weeks) from the time of the message being posted (outliers) is small so it makes an appropriate cut-off point for message analysis to make sure that messages have had a chance to receive all replies. The graphs show a different shape, indicating that response time is less predictable than lifespan.

General (-G) groups show a faster response rate and a shorter lifespan than topic-specific (-S) groups for MG and KS ($p<0.05$, unpaired t-test) perhaps indicating that users in general groups have a broad interest and make conversational replies that do not require a task to be solved.

Within topic-specific categories the TR groups have a

[18] http://www.insidefacebook.com/2013/11/28/infographic-what-types-of-images-are-posted-on-facebook

[19] The group categorisation was done independently by the author and 2 postgraduate researchers at the University of Essex. Where there was not consensus on the categorisation (18%), a final decision was made by the author after group discussion.

	Received reply	Replies (median)	Replies (mean)	Response time	Lifespan	Outliers	Active users	Male users	Workrate (median)	Workrate (mean)
TR-G	81.5%	3	4.1	00:28:30	16:26:16	2.3%	28.0%	69.2%	4	20.8
TR-S	71.0%	2	3.2	00:48:57	11:55:09	1.5%	36.5%	63.4%	4	22.4
MG-G	42.7%	0	1.6	00:58:25	10:25:50	1.4%	20.3%	75.1%	3	12.8
MG-S	49.4%	0	1.8	01:59:46	16:39:43	4.0%	32.4%	59.9%	4	14.5
KS-G	50.5%	1	2.8	00:28:29	07:34:21	0.6%	18.4%	59.4%	3	20.9
KS-S	58.5%	1	2.2	01:24:45	18:12:20	3.1%	38.3%	59.0%	4	11.4

Table 3: A table summarising group categories: the proportion of messages that received a reply; the number of replies (median and mean); the response time (median) for the first reply (hh:mm:ss); the lifespan (median) of the thread (hh:mm:ss); the proportion of outlier replies beyond 1092:16:00; active users (a user who has posted a message or reply) per group (mean); proportion of the active users that are male; and the median and mean workrate (messages/replies per active user).

faster response time and shorter lifespan ($p<0.05$, unpaired t-test) as users of these groups anticipate task requests and are primed to submit a reply, especially if it is an opportunity to demonstrate their knowledge. This would be harder to achieve in general groups because the task posted may be outside the knowledge of most users.

Response time and lifespan of messages are influenced by the interface design of Facebook. When messages are first posted they appear on a user's news feed and/or notifications and the group wall. Over time they are replaced with other messages, move down the page until no longer visible and can only be accessed by clicking for older pages. If a message receives a reply it is moved back to the top of the page (termed "bumping").

Messages posted in the TR groups have more replies than the other groups ($p<0.05$, unpaired t-test). This is unsurprising as these groups are used for posting tasks that require a response, unlike the more passive nature of other groups. This makes the TR groups a good candidate for human computation because more users are potentially involved in the solution of the task.

User activity and workrate

User activity was calculated as the proportion of group members that had posted a message or reply from the total membership at the time of the capture.

Topic-specific groups have more active users ($p<0.05$, z-test), another indication that the community of users in these groups are more engaged with the subject matter and may even know each other personally (as specialist research areas tend to be quite small).

The TR groups have more active members who perform at a higher workrate ($p<0.05$, z-test) than the MG groups, supporting the idea that users joining TR groups are more willing to actively participate in problem solving. Users of MG groups may be more passive by simply enjoying the images and videos being shared.

Collaborative systems, where workload is shared without control, frequently see a Zipfian distribution of workload with only a small proportion of the users doing most of the work (Chamberlain, Kruschwitz, and Poesio 2012) and the distribution of workload amongst these groups is similar.

Gender distribution

The gender distribution of the active users of groups shows a distinct male bias in contrast to other types of social network activity, such as gaming (Chamberlain, Kruschwitz, and Poesio 2012), and Facebook generally, which is reported to have more female users.[20] Only 12% of contributors to Wikipedia are female (Glott, Schmidt, and Ghosh 2010), a statistic that prompted significant research into the gender bias in the authorship of the site (Laniado et al. 2012).

It may be that groupsourcing is appealing in the same way as Wikipedia, or perhaps males prefer image-based tasks to word-based problems to solve (Mason and Watts 2009), or even that marine science and SCUBA diving is a male dominated interest (66% of PADI diving certifications in 2010 were for men).[21]

The different homogeneity of groups can have an impact on their collective intelligence (Woolley et al. 2010) so it is worth noting that TR groups do not have the highest male bias.

Task Distribution and Accuracy

In order to assess the quality of data that could be extracted and to investigate the distribution of the tasks within the group categories 200 threads were selected at random from each category to form a subcorpus of 1,200 threads.

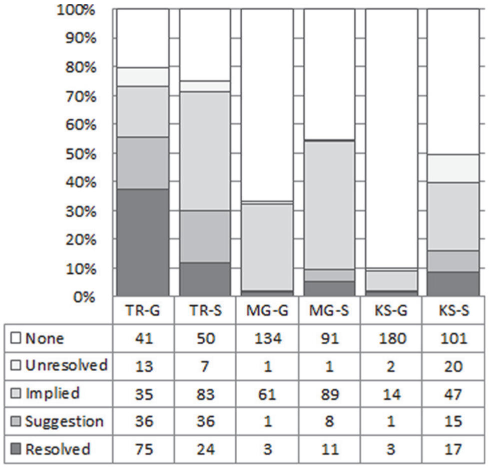
Distribution of task requests

The subcorpus was manually categorised in a random order for data load and inquisition (see *Threads, messages and replies*) by only viewing the thread text and author names, thus each thread could be assigned a type (see Table 1).

Implied, Suggestion and Resolved tasks all contain data that could be extracted to solve the image classification tasks. TR groups have more data loaded threads than MG or KS groups ($p<0.05$, z-test) and it is not surprising due to the purpose of the groups (see Figure 5). Additionally, tasks are more likely to be solved in the TR groups comparing resolved tasks to unresolved tasks ($p<0.05$, z-test).

[20] http://royal.pingdom.com/2009/11/27/study-males-vs-females-in-social-networks

[21] http://www.padi.com/scuba-diving/about-padi/statistics/pdf



	TR-G	TR-S	MG-G	MG-S	KS-G	KS-S
□ None	41	50	134	91	180	101
□ Unresolved	13	7	1	1	2	20
□ Implied	35	83	61	89	14	47
▨ Suggestion	36	36	1	8	1	15
■ Resolved	75	24	3	11	3	17

Figure 5: Distribution of tasks by group category.

Accuracy of task solutions

Based on the previous findings it could be expected that the highest frequency of task requests and more accurate solutions would be found in the TR-S groups although there are fewer explicit tasks compared to TR-G. A single topic-specific area of Opistobranchia (sea slugs or nudibranchs) was chosen in order to check the accuracy of image classification. In this class of animals external morphology is often sufficient to confirm a classification from an image (unlike, for example, sponges) and this is also an active area on social media.

A random sample of threads from 2 groups (Nudibase[22] and NE Atlantic Nudibranchs[23]) from the TR-S subcorpus was taken. Only photo threads were selected and further threads removed if they were unsuitable for the image classification task (for example, not an Opistobranch, multiple species in an image, close-ups, words printed in the image, continuation and/or gallery threads).

Each thread was manually analysed to extract every named entity (or solution to the image classification task) which were normalised to a marine species ontology.[24] Additionally, the thread sentiment was recorded for each named entity including positive and negative opinions and how many people liked the post. Opinions from the same person were normalised but likes were recorded as totals.

The highest rated named entity for an image (totalling messages, replies and likes to replies) was then presented at random with the associated image and checked using a variety of resources including identification websites[25] [26] [27], Wikipedia, Encyclopedia of Life[28], an Android app[29] and books (Picton and Morrow 1994; Debelius and Peyer 2004) relevant to the geographical range of the group. Synonyms were also checked when it was difficult to find a match. Genus (more general than species level) classifications were ignored because the process of classifying an image to this level would be different, involving feature identification across morphological variations.

Crowdsourcing method	Accuracy
Groupsourcing (test set)	0.93
Crowdflower (training) @ $0.05 n=10	0.91
Crowdflower (test set) @ $0.05 n=10	0.49
Happy Match (moths)	0.81
Crowdflower cancer images @ $0.05 n=5	0.72
Expert + Crowdflower cancer images n=5	0.97

Table 4: Comparison of image classification accuracy between different crowdsourcing methods.

The classification was considered correct if the image was confirmed by the majority of the resources with the species name. The classification was not marked if it could not be found in any of the resources (as it could be a new name not updated to the resources) or if there was a split vote between the top rated answer. In total 61 threads were manually analysed using this method (called the test set).

The results show high accuracy for the image classification task (0.93). This represents the upper limit of what could be expected from groupsourcing as other categories of groups may have lower performance. Additionally, there were very few negative statements (0.14 mean negative statements per thread).

Comparison to other crowdsourcing methods

The images from the subcorpus were also classified using Crowdflower[30] to compare the accuracy. Crowdflower users were presented with an image and asked to provide a species name. Web resources were mentioned in the instructions, as well as the requirement for accurate spelling although minor capitalisation mistakes and synonyms were allowed. The configuration selected the top 36% of users on the system to work on the task who were offered $0.05 per image annotated, with 10 answers required for each image.

A training set of 20 images with known answers was created with the most common sea slugs found on the photo sharing website Flickr[31] and high performing users (according to Crowdflower's assessment) were presented with the test set of images.

Results show that whilst there was high accuracy in the training set, the test set showed much lower accuracy (see Table 4). This is an indication of how hard the task was in the test set, as there would be no need to request a classification for a common image. If task difficulty is extrapolated to groupsourcing it would achieve an accuracy of 0.99 on the training data set.

[22] https://www.facebook.com/groups/206426176075326

[23] https://www.facebook.com/groups/NE.Atlantic.nudibranchs

[24] http://www.marinespecies.org

[25] http://www.seaslugforum.net (Bill Rudman)

[26] http://www.nudibranch.org (Jim Anderson)

[27] http://www.medslugs.de (Erwin Köhler)

[28] http://www.eol.org

[29] http://www.inudibranch.com (Gary Cobb)

[30] http://www.crowdflower.com

[31] https://www.flickr.com

Groupsourcing shows higher accuracy compared to a GWAP approach to classifying biological images (Happy Match, moths) (Prestopnik, Crowston, and Wang 2014).

A classification task using images of breast cancer showed reasonable accuracy from Crowdflower using a similar configuration, however an additional approach was to "crowd-power experts" by using crowdsourcing to deal with majority of the easy work and get experts to focus on the difficult images (Eickhoff 2014). This accuracy is comparable to what could be achieved by groupsourcing and could be considered a similar scenario where the majority of group users take on the bulk of the work solving easy tasks leaving the experts to focus on what is of most interest to them. However, the distinction between experts and non-experts in the crowd may not be clear cut (Brabham 2012).

Discussion

This investigation of groupsourcing shows it to be a potentially useful way to complete tasks and perform data collection but can this method be applied to other tasks and other social networking platforms? There have been examples of other tasks being completed on different social networks such as expert finding, job hunting, computer software bug fixing, etc., and these, like the image classification task examined here, are complex human computation tasks that are performed with the collective intelligence of a group. This is unlike the approach of crowdsourcing generally where complex tasks are broken down into smaller chunks that can easily be completed by non-experts. It takes a degree of creativity to imagine mundane tasks in a format that might be applicable to groupsourcing and this may be its biggest disadvantage. Experiments with a text analysis task using groupsourcing have not been as productive because they do not follow exactly the same paradigm. It would also be worth considering other factors when assessing which crowdsourcing approach might be suitable for a different task such as how long it takes to get an answer (wait time), how many tasks can be completed per hour (throughput) and financial costs of setup, advertising and maintenance.

Automatic processing

A significant challenge for groupsourcing as a methodology is the automatic processing of the threads. There is a large quantity of data associated with threads and removing this overhead is essential when processing on a large scale. The natural language processing needs to cope with ill-formed grammar and spelling, and sentences where only context could make sense of the meaning, for example (taken from the subcorpus):

> "And my current puzzle ..."
> "Need assistance with this tunicate please."
> "couldn't find an ID based on these colours"

Users who post requests for solutions to tasks may get better answers if they create a well-formed question and provide as much meta-data as possible, as the lack of both is often a cause of frustration in TR groups.

Aggregating data

For this investigation each thread was manually marked up for named entities and sentiment with majority voting applied to get the best answer from the users involved in the thread. Sophisticated crowd aggregation techniques (Raykar et al. 2010) can be used to gauge the confidence of data extracted from threads on a large scale.

Social networks discourage the expression of negative views of other users' posts and it seems intuitive that positive behaviour be reinforced in crowdsourcing to encourage participation. Facebook has resisted repeated calls from users to add a "dislike" button for presumably this reason, especially as the content is linked to advertising. The low frequency of negative statements found in the test set also suggests that correcting a user's opinion is a socially uncomfortable thing to do, even if it would improve the quality of the solution.

Limitations of social networks

One of the most serious drawbacks of the groupsourcing approach is the changing technology and popularity of social media. Although fairly mature with a high take-up rate, it is still an emerging technology, and changes are made to the terms of service, access and software language that could swiftly render a dependent system redundant.

Another drawback to using social networks is that people use them in different ways and there is no "right" way. There are also a proportion of user accounts used for spreading advertising or for spamming. Users have different expectations that may lead to segregation in groups and data not being entered in a fashion that is expected. Users can also change a post after it has received replies, meaning a user can make a task request and then change the message once a solution has been offered, even deleting replies from the thread dialogue. This is not malicious or ungrateful behaviour but simply a different way of using groups to organise data.

Conclusion

In comparison to other methods of crowdsourcing for image classification, groupsourcing offers a high accuracy, data-driven and low cost approach. Users are self-organised and intrinsically motivated to participate, with open access to the data. By archiving social network data it can be preserved, categorised and explored in meaningful ways. There are significant challenges to automatically process and aggregate data generated from groupsourcing however this study shows the potential for this type of human computation.

Future Work

The method of data caching described here only creates a snapshot of a group. Further development would be required to incorporate the temporal dynamics of social networks and filtering of messages would be required to minimise the database load (Maynard, Bontcheva, and Rout 2012).

Natural language processing would be required for this method to be scalable including extracting more information from the thread dialogue such as features, quantity, location, etc. One such application in the area of marine biology would be the development of groupsourced matrix keys.

Acknowledgments

The author would like to thank Udo Kruschwitz, Massimo Poesio, Dave Smith, Bernard Picton, David Kipling, Kate Starbird and the anonymous reviewers for their comments and input to the research.

References

Agichtein, E.; Castillo, C.; Donato, D.; Gionis, A.; and Mishne, G. 2008. Finding high-quality content in social media. In *Proceedings of WSDM'08*, 183–194.

Brabham, D. C. 2012. The myth of amateur crowds. *Information, Communication and Society* 15(3):394–410.

Bunt, H.; Alexandersson, J.; Choe, J.-W.; Fang, A. C.; Hasida, K.; Petukhova, V.; Popescu-Belis, A.; and Traum, D. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of LREC'12*.

Chamberlain, J.; Fort, K.; Kruschwitz, U.; Mathieu, L.; and Poesio, M. 2013. Using games to create language resources: Successes and limitations of the approach. In *ACM Transactions on Interactive Intelligent Systems*, volume The People's Web Meets NLP: Collaboratively Constructed Language Resources. Springer.

Chamberlain, J.; Kruschwitz, U.; and Poesio, M. 2012. Motivations for participation in socially networked collective intelligence systems. In *Proceedings of CI'12*.

Debelius, H., and Peyer, B. 2004. *Nudibranchs and Sea Snails: Indo-Pacific Field Guide*. IKAN-Unterwasserarchiv.

Eickhoff, C. 2014. Crowd-powered experts: Helping surgeons interpret breast cancer images. In *Proceedings of GamifIR'14*.

Faridani, S.; Bitton, E.; Ryokai, K.; and Goldberg, K. 2010. OpinionSpace: A scalable tool for browsing online comments. In *Proceedings of CHI'10*.

Forsyth, D. 2005. *Group Dynamics*. International student edition. Cengage Learning.

Fort, K.; Adda, G.; and Cohen, K. B. 2011. Amazon Mechanical Turk: Gold mine or coal mine? *Computational Linguistics (editorial)* 37:413–420.

Gao, H.; Wang, X.; Barbier, G.; and Liu, H. 2011. Promoting coordination for disaster relief – From crowdsourcing to coordination. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, volume 6589 of *Lecture Notes in Computer Science*. Springer. 197–204.

Glott, R.; Schmidt, P.; and Ghosh, R. 2010. Wikipedia survey – Overview of results. *UNU-MERIT* 1–11.

He, J.; van Ossenbruggen, J.; and de Vries, A. P. 2013. Do you need experts in the crowd?: A case study in image annotation for marine biology. In *Proceedings of OAIR'13*, 57–60.

Holt, B. G.; Rioja-Nieto, R.; MacNeil, A. M.; Lupton, J.; and Rahbek, C. 2013. Comparing diversity data collected using a protocol designed for volunteers with results from a professional alternative. *Methods in Ecology and Evolution* 4(4):383–392.

Howe, J. 2008. *Crowdsourcing: Why the power of the crowd is driving the future of business*. Crown Publishing Group.

Kittur, A.; Chi, E. H.; and Suh, B. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of CHI'08*, 453–456.

Laniado, D.; Castillo, C.; Kaltenbrunner, A.; and Fuster-Morell, M. 2012. Emotions and dialogue in a peer-production community: The case of Wikipedia. In *Proceedings of WikiSym'12*.

Lu, D., and Weng, Q. 2007. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* 28(5):823–870.

Malone, T.; Laubacher, R.; and Dellarocas, C. 2009. Harnessing crowds: Mapping the genome of collective intelligence. Research Paper No. 4732-09, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA.

Mason, W., and Watts, D. J. 2009. Financial incentives and the "performance of crowds". In *Proceedings of KDD workshop HCOMP'09*.

Maynard, D.; Bontcheva, K.; and Rout, D. 2012. Challenges in developing opinion mining tools for social media. *Proceedings of LREC'12 workshop @NLP can u tag #user_generated_content*.

Nov, O. 2007. What motivates Wikipedians? *Communications of the ACM* 50(11):60–64.

Picton, B., and Morrow, C. 1994. *A field guide to the nudibranchs of the British Isles*. Immel Publishing.

Prestopnik, N.; Crowston, K.; and Wang, J. 2014. Exploring data quality in games with a purpose. In *Proceedings of iConference'14*.

Raykar, V. C.; Yu, S.; Zhao, L. H.; Valadez, G. H.; Florin, C.; Bogoni, L.; and Moy, L. 2010. Learning from crowds. *Journal of Machine Learning Research* 11:1297–1322.

Sidlauskas, B.; Bernard, C.; Bloom, D.; Bronaugh, W.; Clementson, M.; and Vari, R. P. 2011. Ichthyologists hooked on Facebook. *Science* 332(6029):537.

Snow, R.; O'Connor, B.; Jurafsky, D.; and Ng, A. Y. 2008. Cheap and fast - but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of EMNLP'08*.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *Proceedings of CHI'04*.

von Ahn, L.; Maurer, B.; McMillen, C.; Abraham, D.; and Blum, M. 2008. reCAPTCHA: Human-based character recognition via web security measures. *Science* 321(5895):1465–1468.

von Ahn, L. 2006. Games with a purpose. *Computer* 39(6):92–94.

Woolley, A. W.; Chabris, C. F.; Pentland, A.; Hashmi, N.; and Malone, T. W. 2010. Evidence for a collective intelligence factor in the performance of human groups. *Science* 330:686–688.