Harnessing the Intelligence of the Crowd for Problem Solving and Knowledge Discovery

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

Abstract

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. Some approaches are required to incentivise the participation of users, other approaches have self-motivated users but present different challenges. This research investigates the user motivations behind crowdsourcing and human computation by looking at microworking, games-with-a-purpose and groupsourcing approaches in terms of quality, quantity and other factors that determine their utility.
Keywords: games-with-a-purpose, social networks, crowd intelligence

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.

The motivation of the PhD research comes from developing and analysing data from a crowdsourced text annotation game. Whilst the game was comparatively successful in terms of quantity and quality of data it still suffered numerous bottlenecks in data collection. In order to address these issues a second version of the game was deployed on Facebook to make use of the existing network of users and this achieved better results.

Based on these findings the PhD research focused on areas where users were already doing tasks using social networks without a central structure. By applying crowdsourcing methods to aggregate the data more meaningful visualisations and methods of knowledge discovery could be achieved.

Related Work

Three common 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).

These approaches can be seen as examples of the broad term **collective intelligence** (Malone, Laubacher, and Dellarocas 2009).

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. **Citizen science**, where members of the public contribute knowledge to scientific endeavours, is an established predecessor of crowdsourcing and social networks have been successfully used to connect professional scientists with amateur enthusiasts (Sidlauskas et al. 2011). Social networks are self-organized and decentralized; tasks are

created by the users, so they are intrinsically motivated to participate, and the natural language of the interface allows them to express their emotions and frustrations whilst solving tasks.

Groupsourcing, where a task is completed using a group of intrinsically motivated people of varying expertise connected through a social network, is an effective method of human computation in some domains (Chamberlain 2014). Users solve problems with high accuracy, educate each other and share novel information and ideas. Contribution to science, learning and discovery are the driving motivations behind citizen science participation (Raddick et al. 2013).

Research

This PhD research looks at entity classification in text documents (Wikipedia articles and fiction from Project Gutenberg[1]) and images from social networks in the marine biology domain.

In the case of text documents entity classification is through anaphora resolution, the semantic task concerned with recognizing that, e.g., the pronoun *it* and the definite nominal *the town* refer to the same entity as the proper name *Wivenhoe*, and to a different entity from the mentions *Colchester* or *River Colne*.

> Wivenhoe developed as a port and until the late 19th century was effectively a port for Colchester, as large ships were unable to navigate any further up the River Colne, and had two prosperous shipyards. It became an important port for trade for Colchester and developed shipbuilding and fishing industries. The period of greatest prosperity for the town came with the arrival of the railway in 1863.[2]

Anaphora resolution is a key semantic task both from a linguistic perspective and for applications ranging from summarisation to text mining.

Entity recognition is also performed in image classification (where objects in an image are identified). In this case the annotations are open (can be any text) and apply to the whole image. Region annotation, where parts of an image are annotated, is more complex.

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[3] and most recently seen with the disappearance of Malaysia Airlines flight MH370 in 2014.[4] 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)[5] [6] [7], whilst others require the users to supply the source data as well as the classification.[8] [9] [10] 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).

In both types of task an occurrence of an entity is identified that can be used to build an entity concept. This research investigates several themes that are essential for harnessing the intelligence of the crowd to complete this task.

User participation

One of the most significant failings of human computation systems is the lack of participation from users. Incentives are commonly divided into personal, social and financial categories but what is most the most effective motivator for a particular task?

Task design

The interface design needs to be appropriate for the task, the intended users and level of difficulty but how should data quantity and quality be balanced? Is it better to collect more noisy data or less higher quality data? How will the task difficulty affect the users? Will they rise to the challenge of difficult tasks or be put off because the incentives are not sufficient?

Assessing users and data

Assessing users and the data they contribute is a key part of this research. Attention slips, malicious input and poorly trained users need to be differentiated from genuinely ambiguous data and this can be done at the point of data entry or in post-processing. Without a known set of answers to judge users by it may be necessary to rely on user reaction time (Chamberlain and O'Reilly 2014) or modelling their behaviour over time (Passonneau and Carpenter 2013).

Aggregating data

Once the data has been collected it needs to be aggregated to produce a resource with a set of answers to the tasks. This data can then be compared with traditional methods of completing the tasks to assess the utility of the crowdsourcing approaches in terms of time, complexity and financial investment.

[1] http://www.gutenberg.org

[2] Taken from Wikipedia's page about Wivenhoe, the village next to the University of Essex.

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

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

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

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

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

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

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

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

Figure 1: Screenshot of Phrase Detectives on Facebook.

Methodology

The PhD research investigates 3 approaches for problem solving and knowledge discovery:

Text annotation with a game

Phrase Detectives[11] is a single-player game-with-a-purpose developed to collect data about anaphora and is centred around a detective metaphor (Chamberlain, Poesio, and Kruschwitz 2008). The game architecture is articulated around a number of tasks and uses scoring, progression and a variety of other mechanisms to make the activity enjoyable. A mixture of incentives, from the personal (scoring, levels) to the social (competing for some players, participating in a worthwhile enterprise for others) to the financial (small prizes) are employed. This approach was adopted not just to annotate large amounts of text, but also to collect a large number of judgements about each linguistic expression.

The Facebook version of Phrase Detectives[12], launched in February 2011, maintained the overall game architecture whilst incorporating a number of new features developed specifically for the social network platform. The game was developed in PHP SDK (a Facebook API language allowing access to user data, friend lists, wall posting etc) and integrates seamlessly within the Facebook site.

The game uses 2 styles of text annotation for players to complete a linguistic task. Initially text is presented in Annotation Mode (called Name the Culprit in the game). This is a straightforward annotation mode where the player makes an annotation decision about a highlighted markable (section of text). If different players enter different interpretations for a markable then each interpretation is presented to more players in Validation Mode (called Detectives Conference in the game). The players in Validation Mode have to agree or disagree with the interpretation.

[11] https://www.phrasedetectives.com

[12] https://apps.facebook.com/phrasedetectives

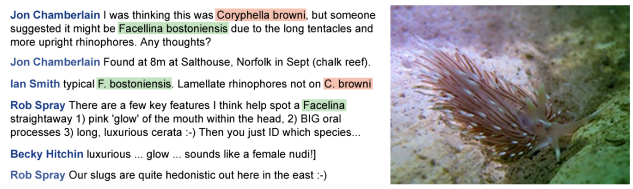

Figure 2: Detail of a typical message from Facebook containing an image classification task having been analysed for named entities.

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

Table 1: Comparison of image classification accuracy between groupsourcing and microworking.

Image classification with groupsourcing

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.[13]

The accuracy of the image tags has been shown to be very high in some domains (Chamberlain 2014) however automatically aggregating the data is not trivial. Ontologies, gazetteers or controlled vocabularies can be used to structure the content. This research uses, in the first instance, an ontology of marine species[14] as a hierarchical list of named entities. Each chunk of text from a message thread is scanned for named entities from the ontology and an index table is created in a MySQL database - see Figure 2.

The named entity index is used to aggregate the messages allowing a user to find all content regarding a particular marine species and other species that are associated with it (what it eats, what it looks similar to, etc.). Additionally, messages containing a named entity with an image attached are used to create a gallery of photographic examples of the species.

Microworking on Crowdflower

There has been considerable research into user behaviour on Amazon's Mechanical Turk and Crowdflower and it has become a default research tool for collecting small quantities of research data. Text and image data from the above approaches have been annotated using Crowdflower in order to benchmark the task difficulty using the same interface for different media (see Table 1 for a comparison of Crowdflower with Groupsourcing).

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

[14] http://www.marinespecies.org (Sept 2012)

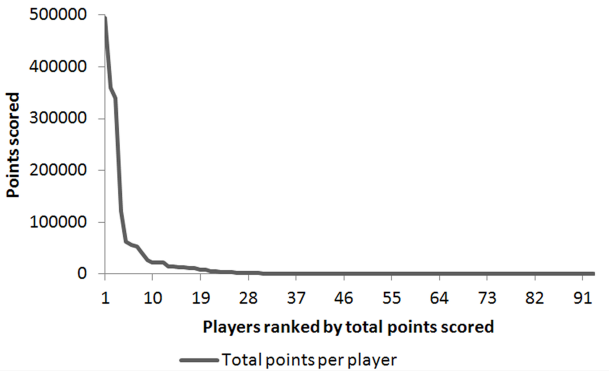

Figure 3: Chart showing the ranked scores of players from Phrase Detectives on Facebook.

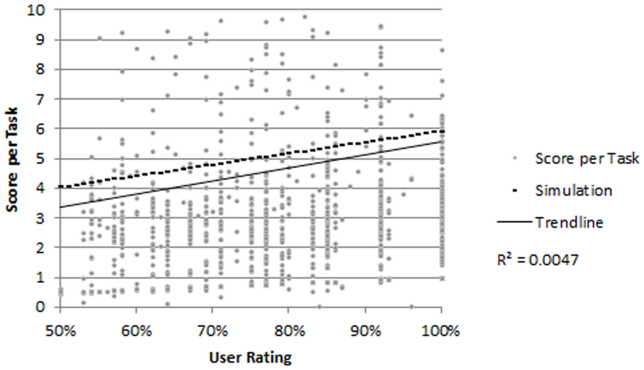

Figure 4: Actual score per task from Phrase Detectives compared to a simulation model based on task difficulty, corpora completion and average ability of the crowd.

Challenges

There are numerous challenges in this area of research.

Application of incentives

It is increasingly difficult to attract the interest of users however the principle of using personal, social or financial incentives seem to still apply. New or adapted strategies are some of the most discussed topics amongst the developers of human computation systems.

Workload distribution

Studies of user contribution in Phrase Detectives show that the ten highest scoring players (representing 1.3% of total players) had 60% of the total points on the system and had made 73% of the annotations (Chamberlain, Poesio, and Kruschwitz 2009). In the Facebook version of the game the ten highest scoring players (representing 1.6% of total players) had 89% of the total points and had made 89% of the annotations (see Figure 3). A similar Zipfian distribution of workload is seen with the users of groupsourcing and other approaches.

These results show that the majority of the workload is being done by a handful of users. However, the influence of users who only contribute a little should not be undervalued as in some systems it can be as high as 30% of the workload (Kanefsky, Barlow, and Gulick 2001) and this is what makes the collective decision making robust.

Crowd homogeneity

The gender distribution of the active users of groupsourcing shows a distinct male bias in contrast to other types of social network gaming (Chamberlain, Kruschwitz, and Poesio 2012), and Facebook generally, which is reported to have more female users.[15] 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).

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

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 the topic investigated is a male dominated interest. The different homogeneity can have an impact on collective intelligence (Woolley et al. 2010).

Crowd-powered experts

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).

Automatic processing

The pre-processing of text used for Phrase Detectives was reasonably accurate but there were errors that had to be manually corrected that, given the size of the corpus, took considerable time. A significant challenge for groupsourcing as a methodology is the automatic processing 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 previous context could make sense of the meaning, for example:

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

Aggregating data

Currently all the methodologies used in the research use a majority voting aggregation to produce the best answer (see

Figure 4) however sophisticated crowd aggregation techniques (Raykar et al. 2010) could be used to gauge the confidence of data extracted from threads on a large scale.

References

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

Chamberlain, J., and O'Reilly, C. 2014. User performance indicators in task-based data collection systems. In *Proceedings of MindTheGap'14*.

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*.

Chamberlain, J.; Poesio, M.; and Kruschwitz, U. 2008. Phrase Detectives: A web-based collaborative annotation game. In *Proceedings of the International Conference on Semantic Systems (I-Semantics'08)*.

Chamberlain, J.; Poesio, M.; and Kruschwitz, U. 2009. A new life for a dead parrot: Incentive structures in the Phrase Detectives game. In *Proceedings of the WWW 2009 Workshop on Web Incentives (WEBCENTIVES'09)*.

Chamberlain, J. 2014. Groupsourcing: Distributed problem solving using social networks. In *Proceedings of HCOMP14*.

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

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

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.

Kanefsky, B.; Barlow, N.; and Gulick, V. 2001. Can distributed volunteers accomplish massive data analysis tasks? *Lunar and Planetary Science* XXXII.

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*.

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*.

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

Passonneau, R. J., and Carpenter, B. 2013. The benefits of a model of annotation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, 187–195. Association for Computational Linguistics.

Raddick, M. J.; Bracey, G.; Gay, P. L.; Lintott, C. J.; Cardamone, C.; Murray, P.; Schawinski, K.; Szalay, A. S.; and Vandenberg, J. 2013. Galaxy Zoo: Motivations of citizen scientists. *ArXiv e-prints*.

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.; 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.