Doing Useful Work Using Games

Star, K

Original citation & hyperlink:
Star, K 2014, Doing Useful Work Using Games. in AD Gloria (ed.), Lecture Notes in Computer Science. vol. 8605, Springer International Publishing, Switzerland, pp. 316-323. DOI: 10.1007/978-3-319-12157-4_25
https://dx.doi.org/10.1007/978-3-319-12157-4_25

DOI 10.1007/978-3-319-12157-4_25
ISSN 0302-9743

Publisher: Springer

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-12157-4_25

Doing useful work using Games - a brief review
Kam Star – January 2013

" In every job that must be done, there is an element of fun. You find the fun, and - SNAP - the job's a game! " (Mary Poppins, 1964 [FILM])

Abstract

This is a short review of four papers on the generation of useful metadata and solving scientific problems through play. The central thesis of the papers reviewed conclude that games that are fun to play may also be used to carry out useful activity, for example in solving scientific questions. Three of the papers primarily demonstrate use of games for carrying out tasks that would be difficult for automated algorithms. The final paper demonstrates the use of a game as a learning tool with the ability to generate useful metadata as a side benefit. Throughout the review, design elements which have demonstrated contribution to the effectiveness of the games are highlighted. Additionally an objective method for evaluating the performance of the games is briefly discussed.

Each paper discusses one or more games. For the sake of clarity one game is picked from each paper. The papers selected are :

> *Designing Games With A Purpose.* (Von Ahn and Dabbish : 2008) Focusing on the **ESP game** aka Google Image Labeler[1], which provides meaningful tags for images on the Web, a task which is almost impossible to achieve using computer-vision algorithms.
>
> *The challenge of designing scientific discovery games.* (Cooper et al. 2010) - Focusing on **Foldit**[2]. A biochemical discovery game. Designed to make it possible for non-experts to make useful contribution to the scientific domain of protein folding.
>
> *Validation of Music Metadata via Game with a Purpose.* (Dulavcka and Bieliková : 2012) Focusing on **CityLights**[3]. A music metadata validation game.
>
> *The Spectral Game : leveraging Open Data and crowdsourcing for education.* (Bradley et al. 2009) Focusing on **Spectral game**[4], a game that involves matching molecules to their spectra graph.

[1] ESP be played at http://images.google.com/imagelabeler/
[2] Foldit be played at http://fold.it
[3] CityLights be played at http://bit.ly/city-lights
[4] Spectral Game be played at http://www.spectralgame.com/

Introduction : Calculations that are easy for humans but hard for algorithms

According to researcher at the Entertainment Software Association, by the age of 21 the average American spends the equivalent of five years of working a full time job 40 hours a week playing computer games.[5] The constructive channelling of human brainpower through computer games may potentially be an untapped source of intellectual capacity. Arvidsson and Sandvik (2007) argue that playing games may be regarded as a kind of unpaid immaterial labour, implying players' creativity and general intellect may be utilised for the benefit of the game designer's intent.

Some trivial tasks which are easily solved by people can challenge the most sophisticated algorithms and artificial intelligence. Computer games which utilise human intelligence in order to carry out useful tasks otherwise difficult for algorithms to perform are described as 'games with a purpose' or **GWAP** by Von Ahn and Dabbish (2008).

Scientific discovery games or **SDG** as described by Cooper et al. (2010) are a type of GWAP that are concerned with harnessing the enormous collective problem-solving potential of the game playing population, but who are not familiar with the specific scientific domain.

SDGs are about using human problem solving ability to solve computationally difficult scientific problems. They provide game like mechanisms for non-expert players to help solve these problems. Foldit fits perfectly within this description. Crowdsourcing the discovery of protein's natural shapes.

The Spectral game is also arguably an SDG. Although its creators see it more as a Serious Game. It's aims is to leverage open data and crowdsourcing for education. The game involves matching molecules to various forms of interactive spectra graphs. Interpretation of spectra is an essential skill for organic chemists and many students struggle to grasp the nuances of various spectroscopy techniques. It was created by bringing together Open Source spectral data, a spectrum viewing tool and a workflow that combined these within a gaming framework.

CityLights is a GWAP and has many similarities with ESP although its aims is to validate the accuracy of existing metadata associated with a song, such as mood, quality, time and place to listen to the song. It was designed to use player competition and music exploration as its principle fun aspects.

Elements for success

[5] Entertainment Software Association - http://www.theesa.com/facts/gamer_data.php

This section examines some of the critical design elements relating to the success of each of the games as described by the authors of the related papers. The frameworks formulating each approach to developing the various games are also identified and briefly discussed.

Von Ahn and Dabbish (2008) argue that the critical ingredient for a successful GWAP lie in persuading enough human-hours to be spent playing the games with a probabilistic guarantee that the game's output is correct. They also present three frameworks for designing and deploying GWAPs, covering basic rules and winning condition that lead to optimum outcomes - that is holding the player's interest whilst compelling them to perform the intended computation. The three types of game have been described are :

A. Output-agreement Games : Where two players are given the same input and must produce the same output. For example input is an image and the output would be a matching keyword as in ESP.

B. Inversion-problem games : Where one player is the "describer" and the other player the "guesser". To win the guesser must give the same output that was given to the describer.

C. Input-agreement games : Where two players are instructed to produce output describing their input, then determine whether they have been given the same input.

Amos and Kahneman (1981) demonstrate that the way in which a choice is framed, that is the words and phrases used to describe the choice to the decision maker, has a direct relation to its propensity to persuade.

It can be argued that one of the ingredients of success of these games is in the way the game frames the computational task to the player. For instance In the Output-agreement game, rather than asking players to describe what they see, players are asked to "think like the other player", and to describe what the other player may be thinking.

Research confirms humans along with other social animals are wired to be emphatic (Decety and Philips, 2004). In essence people are wired to be able to think like others. Decety and Philips argue that this ability to think like others is an innate human motivation requiring a multitude of cognitive mechanisms. They further report that the degree with which the subject is successful in completing the cycle of empathy displays a strong link with reward mechanisms within the brain. Therefore thinking liking other players may be described as a rewarding social engagement mechanism which would help to increase fun and enjoyment from a game (Decety and Philips, 2004).

Games such as charades and the 2012 game app “Draw Something!” a game where the players play in pairs to guess what the other is drawing. Bought by Zynga for $200M a mere 6 weeks after it was developed [6], rely on this type of mechanism.

The Inversion-problem game is designed in a way to collect facts as a side effect of playing - i.e. the player is never asked to “enter facts about milk” rather the task is framed as “enter facts about milk that will help the other person guess its milk” - the subtle difference between the two approaches means a mundane computational task is transformed from a task about entering the player’s knowledge to a task about emphatic thinking.

In order to increase player enjoyment some additional elements can be incorporated into the Inversion-problem game. These included transparency and alternation. Transparency allows the describer to give small hits to the guesser, making the process much more interactive. Whilst alternation meant that after each round the players swap roles. Creating variety and fully involving both players.

In order to discourage players from random guesses the scoring in input-agreement games strongly penalise incorrect guesses. Rather than de-motivating the player by punishing them with negative points, the point system can be designed to increase points for streaks of correct answers.

The design process for each of the games seems to have been entirely different. With the Foldit team being the only one that included game designers. Foldit brought together biochemistry, game design, art and computer science - in an interactive, multi-disciplinary and iterative approach to design. Where the design was continually evolved in response to gameplay traces, player feedback and expert analysis.

The development of Foldit is ongoing, and according to the authors of the game, the game is constantly being updated. Playtesting is used as a way of uncovering what element of the game are fun and which can be most confusing and difficult to understand. One form of play testing carried out by the team is ‘think-alouds’ these have players play through the levels and say out loud what they are thinking (Ramey et al. 2006). This helps to decipher what works well and what could be adjusted to make the gameplay more fun.

Foldit is essentially a 3D jigsaw puzzle, a simulation tool for shaping proteins framed within a context of a game with challenges and competition. It was developed to show that human’s innate spatial reasoning can make it possible for non-experts to make useful contributions to the field of protein shape problems.(Cooper et al. 2010) In Foldit the solution for the problem is unknown, therefore the game could not be designed with a specific solution in mind. Rather the game must provide the tools needed to assemble the solution, in a similar way as lego provides

[6] http://articles.businessinsider.com/2012-03-21/tech/31218846_1_zynga-revenue-rate accessed 1.1.2013

the building blocks to create anything the player desires. These types of game can be described as open-ended simulation games or sandbox games (Squire : 2007).

According to wikipedia, the definition of sandbox games encompasses a large variety of different types of games, from non-linear entertainment games to virtual worlds and simulation games. Sweetser (2008) defines sandbox games as simulations, except for the somewhat loose definitions of tasks, challenges, and completion.

Squire (2007) identifies the use of sandbox games for their capacity to recruit diverse interests, provide an engaging interface for creative problem solving, and enable players to carry out productive acts. Therefore the use of this type of sandbox framework for a GWAP appears to be well suited to situations where creative problem solving is required. By providing the building blocks and defining the interaction behaviours that mimic the real-world constraints of the problem within the game, the players may be able to discover previously unknown scientific solutions. (Cooper et al. 2010)

Exploring Mechanisms for Player Engagement

The central tenet of GWAPs is player enjoyment while carrying out the computational task, as such players are not directly instructed to solve the computational problem - rather to think like the other player. Other features incorporated in GWAP are taken from challenges outline by Malone (1980, 1982) - these include timed response, score keeping, player skill level, high score lists and randomness.

Timed response is an effective motivator since motivational theory posits that goals that are both well-specified and challenging lead to higher levels of effort than goals that are too vague or easy to accomplish (Locke and Latham : 1990).

The Foldit game uses timed response by having scientists post problems to the server, these puzzles are usually available for a week, before a winner is announced. Similarly on the ESP and Spectra Game time response is used to motivate players for their input.

Player Skill Levels is also used by the ESP and Foldit game. Here players are shown their current skill level and the number of points needed to reach the next level. The ESP game has provided a tremendous amount of data to support the effectiveness of player skill level ranking as it relates to engagement. Data from 200K players indicates that 42% of scores fall within a few points of a rank cut-off. Given that the point intervals for ranking within ESP occupy less than 2% of the possible cumulative scores, the data strongly suggests that many players continue playing just to reach a new rank. The data provided does not indicated whether after levelling up to a new rank and logging off, the players came back to improve their ranking. Goh et al. (2011) argue that displaying the players' skill level to other players is a contributor in replayability of image tagging games.

Scoring in SDGs must direct players toward the solution by encouraging them to explore the solution space. For this Foldit used a competition form where player's goal is to do better than other players. Although players can also collaborate as teams - all players are ranked against each other in terms of their effectiveness in a highscore list.

Highscore lists within EPS were divided into multi-levels, from the last hour to day to all time, varying in difficulty and providing strong, positive motivation for extended play. Particularly useful for those who are competitive.

The authors of CityLights assert that by giving points even for actions not connected with their purpose players will return in the future (Dulavcka and Bielikováet : 2012). They call this "proper scoring" and assert that player enjoyment is provided by highscore lists, however these assertions are not shared by other researchers in this area such as Shell (2010) and Deterding et al.(2011), who argue that for scoring to count, it must be meaningful. This suggests that meaningless scoring for actions that are not related to the skills within the game or as part of a feedback loop is unlikely to be an effective mechanisms for the purpose of replayability.

The gaming elements within the Spectral Game included increasing difficulty as the player progressed through the game by an ever increasing complexity of the spectra. It also included a timed element of finding the right answer before a countdown expires. Players were given their performance relative to recent and top players. Players can also associate themselves with groups which could help in direct score comparison amongst members through the highscore lists.

The output in GWAPs must be accurate and the development team used a variety of techniques including random matching, player testing and repetition - until a certain number of consistent outcomes had been achieved. Von Ahn and Dabbish (2008) additionally used paid participants and independent raters to evaluate the quality of the output produced in the game to ensure it meets the required standards.

In order for a relatively complex game to be better received by players, the players must be introduced to the components of the game in a step by step manner (Salen, Zimmerman : 2003) - within SDGs the gameplay, visualisations, interactions and evaluation methods all require some training - carried out as part of introductory levels to the game. (Cooper et al. 2010)

Accurate visualisation is one of the key aspects of an SDG (Cooper et al. 2010). The visualisation must sufficiently reflect and illuminate the natural rules of the system. In Foldit this meant showing the fundamental properties of proteins. The visualisations must also manage and hide the complexity of the system such that players are not immediately overwhelmed - this was achieved by hiding all but the most critical aspects of the puzzle as the player progressed. Visualisations must also be approachable by players rather than baffle them, for this the Foldit team used a cartoonish look, and enabled players to customize their view as they become more experienced with the system.

Interactions within SDGs must respect the constraints of the system, that is the possible solutions must be plausible within the scientific domain. Interactions must also be sufficient to explore the solution space; it must be possible to achieve the correct output from the given input. Within Foldit this was achieved by running puzzles where the solution was visible as a guide, this demonstrated that it was possible to achieve the correct output with the interactions given to the player.

In a successful GWAP the interactions should be intuitive and fun - within Foldit this concept is brought to life as 'touchability'. The ability to interact with the protein as though you could actually touch it, push it, pull it and shake it. The difference between a GWAP and a general online game (or website) whose aim is to capture screen-time, is in the formers' ability for useful work to be carried out as a by-product of play, whilst the latter is often only concerned with play for fun or in the case of a website, potentially devoid of the play as an engagement mechanism.

Evaluating success

GWAPs can be evaluated in terms of their efficiency and expected contribution. Von Ahn and Dabbish (2008) define the method of measuring the performance of this type of game as; the average number of problem instances solved per human hour, multiplied by the average amount of overall time an individual will play the game.

For instance if during one hour of playing the game an average player carries out 100 tasks and the average lifetime playability of the game per player is 2 hours. The expected contribution per player (CPP) is a score of 200. This CPP score represents an objective measure which can be used to compare and assess the effectiveness of variety of a GWAPs around a specific topic. Since CPP is not concerned with the number of players, rather the expected output from each individual playing the game, the task of marketing and dissemination can be ignored. A game with a higher CPP score is more valuable.

In order to evaluate whether Foldit was capable of being used to find tangible solutions, the solutions were entered into a protein structure prediction methods competition. The results found by players compared favourably to other methods used including automated systems. Cooper et al. (2010) concluded that the game has been designed in such a way that players can use it to solve the specific biochemical problems.

Evaluation of CityLight was carried out by comparison of the outcomes to a priori created by experts. Since only 78 players took part and 50% of the tags did not receive sufficient evaluation. According to Dulavcka and Bieliková (2012) 66% (n=1300) of the tags were evaluated correctly. However in considering the vast number of tags which were not evaluated the game is less than 33%(n=78) accurate in identifying the correct tags, with a failure rate of over 67%(n=78).

Dulavcka and Bieliková (2012) assert that the common problem with most existing GWAPs is a cold-start which causes an insufficient number of players from engaging with the game. However CityLights attracted less than 80 players, compared to all other GWAPs reviewed which many thousands or hundreds of thousands of players. This supports the hypothesis that the cold-start is potentially not the common problem, rather the lack of fun and interest as identified by Von Ahn and Dabbish (2008).

The key feature of the CityLight game was to show tags that had strong support or not so good support. However since tags are displayed to the player grouped together the value of the validation approach is limited because a correct tag may be mixed in a group with incorrect tags and the player would have no choice but to label the entire group incorrect.

Although the definition of the Spectral Game is more fitting with a Serious Game because of its primary goal as an educational tool rather than a human computation output. The authors have noted one of the critical side benefits of the game has been the examination of the open data and reporting of potential issues with it. Since players can flag incorrect spectra and leave comments associated with it for the curators of the open data. This has resulted in the deletion, reassociation and correction of certain spectra from the database through the crowd effort.

The methods of objective evaluation described by Von Ahn and Dabbish (2008) was unique amongst all the paper reviewed and deserves to be adopted by others as a scientific approach to measuring the effectiveness of a GWAP and improvement in game design.

Other Related Work

Using people to solve discrete tasks that computers find difficult to accomplish automatically is not new - Amazon's Mechanical Turk (AMT) (developed in 2005, www.mturk.com) is a platform within which individuals carry out small tasks for small sums of money. The tasks are knows as HITs (Human Intelligence Tasks), and include activities such as choosing the best among several photographs of a storefront, writing product descriptions, or identifying performers on music CDs.

The difference between AMT and GWAPs are in the lack of financial incentives insofar as people willingly perform HITs in GWAPs without a need for monetary remuneration.

The Open Mind Initiative is a worldwide research programme aimed at teaching computer programs commonsense facts - whilst OMI shares many of the elements of GWAPs, the difference in the way that GWAPs are designed to be enjoyable.

Conclusion

The papers reviewed here make a compelling case in the argument for the use of games as a method of carrying out useful computational work by players. The studies examined suggest that is it possible to carry out useful work as a side benefit of playing these games. Moreover that this is a type of work that could be paid for, being carried out for free by the players. In effect free immaterial labour in return for fun.

Some of the traits making GWAPs successful were identified as : Framing, competition, timed-response, transparency, alternation, streak-scoring, high-scores, step-by-step introduction, increasing difficulty and appropriate visualisation and manipulation tools within the game.

As a method of objectively measuring the effectiveness of various GWAPs, a CPP (Calculations Per Player) score has been identified. CPP represents the total number of calculations a single player will on average perform during a lifetime play of a particular game.

Further research into the traits and design elements that contribute positively to the success of GWAPs, building on the work reviewed here, could perhaps yield a ‘best practice’ or guidelines for designers and developers of this type of serious game.

Bibliography

Arvidsson, A., Sandvik, K., 2007. Gameplay as design: uses of computer players' immaterial labour. Northern Lights: Film and Media Studies Yearbook 5, 89–104.

Bradley, J.C., Lancashire, R.J., Lang, A.S.I.D., Williams, A.J., 2009. The Spectral Game: leveraging Open Data and crowdsourcing for education. Journal of cheminformatics 1, 1–10.

Cooper, S., Treuille, A., Barbero, J., Leaver-Fay, A., Tuite, K., Khatib, F., Snyder, A.C., Beenen, M., Salesin, D., Baker, D., 2010. The challenge of designing scientific discovery games, in: Proceedings of the Fifth International Conference on the Foundations of Digital Games. pp. 40–47.

Decety, J., Jackson, P.L., 2004. The functional architecture of human empathy. Behavioral and cognitive neuroscience reviews 3, 71–100.

Deterding, S., Dixon, D., Khaled, R., Nacke, L., 2011. From game design elements to gamefulness: defining gamification, in: Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments. pp. 9–15.

Dula\vcka, P., Bieliková, M., 2012. Validation of music metadata via game with a purpose, in: Proceedings of the 8th International Conference on Semantic Systems. pp. 177–180.

Goh, D.H.L., Ang, R.P., Lee, C.S., Chua, A.Y.K., 2011. Fight or unite: Investigating game genres for image tagging. Journal of the American Society for Information Science and Technology 62, 1311–1324.

Jesse Schell: When games invade real life | Video on TED.com, n.d. .

Locke, E.A., Frederick, E., Lee, C., Bobko, P., 1984. Effect of self-efficacy, goals, and task strategies on task performance. Journal of applied psychology 69, 241.

Locke, E.A., Latham, G.P., 1990. A theory of goal setting & task performance. Prentice-Hall, Inc.

Malone, T.W., 1980. What makes things fun to learn? Heuristics for designing instructional computer games.

Malone, T.W., 1982. Heuristics for designing enjoyable user interfaces: Lessons from computer games, in: Proceedings of the 1982 Conference on Human Factors in Computing Systems. pp. 63–68.

Open world, 2012. . Wikipedia, the free encyclopedia.

Ramey, J., Boren, T., Cuddihy, E., Dumas, J., Guan, Z., Van den Haak, M.J., De Jong, M.D.T., 2006. Does think aloud work?: how do we know?, in: CHI'06 Extended Abstracts on Human Factors in Computing Systems. pp. 45–48.

Salen, K., Zimmerman, E., 2003. Rules of play: Game design fundamentals. MIT press.

Shneiderman, B., 2004. Designing for fun: how can we design user interfaces to be more fun? interactions 11, 48–50.

Squire, K., 2007. Open-ended video games: A model for developing learning for the interactive age. The John D. and Catherine T. MacArthur Foundation series on digital media and learning 167–198.

Sweetser, P., 2008. Emergence in games.

Tversky, A., Kahneman, D., 1981. The framing of decisions and the psychology of choice. Science 211, 453–458.

Von Ahn, L., Dabbish, L., 2008. Designing games with a purpose. Communications of the ACM 51, 58–67.

Yarow, J., n.d. Here's Why $200 Million Is Cheap For Draw Something [WWW Document]. Business Insider. URL http://articles.businessinsider.com/2012-03-21/tech/31218846_1_zynga-revenue-rate (accessed 1.1.13).