Proceedings of the Federated Conference on Computer Science and Information Systems pp. 405–409
DOI: 10.15439/2016F172
ACSIS, Vol. 8. ISSN 2300-5963

Game with a Purpose for Mappings Verification

Tomasz Boiński
Department of Computer Architecture
Faculty of Electronics, Telecommunication and Informatics
Gdańsk University of Technology
11/12 Narutowicza Street
80-233 Gdańsk, Poland
Email: tobo@eti.pg.gda.pl

***Abstract*—Mappings verification is a laborious task. The paper presents a Game with a Purpose based system for verification of automatically generated mappings. General description of idea standing behind the games with the purpose is given. Description of TGame system, a 2D platform mobile game with verification process included in the gameplay, is provided. Additional mechanisms for anti-cheating, increasing player's motivation and gathering feedback are also presented. Example of the system usage for verification of mappings between WordNet synsets and Wikipedia articles is presented. The evaluation of proposed solution and future work is also described.**

I. Introduction

NOWADAYS people tend to spend a lot of time playing computer games. Increased availability of powerful mobile devices further increases time spend on this form of entertainment. In 2012 Samsung Electronics Polska performed a study among people in Poland on the time spend on video games [1]. Almost half of the population aged 27-35 spends 1 to 2 hours weekly playing games and 14% spends over 20 hours a week. High percentage of the players use mobile devices like smartphones (20%) and tablets (5%). It can be seen that in many cases playing games occupies the amount of time equal to at least a part-time job. On the other hand many nowadays problems still cannot be solved by a computer algorithm and assigning human resources to perform such tasks is often economically inefficient.

The question arises what if we could use all that potential resources (time and knowledge of the users and hardware capabilities of their devices) to solve also non-algorithmic problems? One can imagine that if we would treat a group of users as a distributed system, then it is sufficient to divide a problem into small portions, distribute them to the players and finally aggregate achieved results. This however introduces some difficulties, from technical ones like how to divide a problem into sub-problems, how to distribute them and how to gather results, to more social oriented like how to trust the results and more importantly how to convince people to spend their time on solving our problem. Some research shows however that even educational games can be well perceived if constructed properly [2] so that an algorithm verification system within games seems plausible.

Numerically solvable problems adopted volunteer computing model [3] where the users donate the power of their machines when it is not needed (the calculations are done between the periods of active hardware usage). Using this model it is not possible to solve all type of problems, as some of them cannot be successfully turned into a computer algorithm [4]. We can use heuristics but than we still have to verify the results manually. In crowdsourcing [5] approach the user is encouraged to perform a task for some type of gratitude. The task can be both algorithmic and non-algorithmic.

It is also worth mentioning that the problems that are difficult for computers are usually quite easy for humans. Example of such problem is image recognition, image tagging, natural language understanding and processing or verification of results obtained by traditional heuristics. The results often very complicated algorithms are quite easy to grasp by an average human. Linking those two areas could prove to be useful for performing laborious tasks without a need of hiring additional workers especially that many research results needs some evaluation and sometimes it is the sole purpose of the research [6].

In this paper we focus on a so called human-based computation (HBC) [7] model. It is using human brain directly to solve a computational problem. The term was defined by Kosorukoff in 2001 in a paper about human enhanced genetic algorithm [8].

HBC can be viewed as similar to crowdsourcing. The later focuses solely on solving the problem by human, while in the former model part of the problem is solved by a computer. Usually the machine performs sub-problem organization, distribution and retrieval of results, sometimes some calculations are done using heuristics. The human part usually contains the verification of computer generated results or performing the calculations itself [9].

In our research we applied HBC-model for verification of mappings. As an example we used mappings between WordNet and Wikipedia [10], [11], [12], [13] that were obtained during Colabmap[1] project and are a result of running heuristics on a computer. Currently we are working on generalization of the proposed solution hoping to provide a general framework suitable for solving different types of problems.

II. Games with a Purpose

In 2006 Luis von Ahn proposed usage of computer games as something more than pure entertainment and thus creating

[1] http://kask.eti.pg.gda.pl/colabmap

978-83-60810-90-3/$25.00

the idea of so called GWAP (Game With A Purpose) [4]. GWAPs are typical games that provide standard entertainment value that users expect but are designed in a way that allows generation of added value by solving a problem requiring intellectual activity. It is worth noticing that GWAPs does not allow financial gratification for the work. The will to continue playing should be treated as the only way of gratifying users [14].

Ahn defined three types of GWAPs:

- output-agreement game,
- inversion-problem game,
- input-agreement game.

In the first type of GWAPs two randomly selected players are presented with identical input data and each produces results only based on the available information. Both players should achieve identical results without any knowledge about the other player - they are awarded only when both will give identical answers. In this case an identical answer provided by both players is treated as highly probable to be correct as it comes from independent sources. Example of such game is The ESP Game [15], where users task was to tag images with keywords. The players were presented with an image and were given 2,5 minute to enter all keywords that are related to the image. The game proved to be very popular. During first few months authors managed to gather around 10 million tags with statistics showing many users playing for over 40 hours a week [4]. In 2006 Google released their own version of the game called Google Image Labeler[2] (it was shut down in 2011) which was used to extend capabilities of Google Graphics.

The second type, the inversion-problem game, also selects players randomly. The players are however divided into two groups - describers and guessers. The describer is presented with input data and is responsible for creating tips allowing the guessing player to correctly point out the input data. The players are awarded points when the output given by the guesser is equal to the input. One of the examples of such game is Phetch [16]. One of the players is presented with a random image from the Internet. His or her task is to describe the image to other players. Other players task is to find an identical image. Other example is the Peekaboom game [17]. The task of the players is to quest words that are describing an image. The "boom" player is presented with an image and its description in a few words. The "peek" player is presented with empty page on which the "boom" player gradually reveals parts of the image. The "peek" player have to guess, based on the revealed fragments, the exact words describing the image.

The third type, input agreement game, also selects players randomly. Both players have to achieve agreement on the input data. More precisely they have to guess whether the other player received the same or different input data. Each player describes what he or she sees on the screen and the other player have to state whether the input is similar to theirs or different. The example of such game is TagATune [18] where players should describe their feelings about a tune that is played. Based on the description the players have to decide whether they heard the same or different tune.

What is common for all above types of games is that the players unknowingly generate added value that is not possible to calculate using computers. The problem behind such games is a way to lure players - only very large user base can provide viable results. During implementation many techniques can be used to enrich the game and encourage more players, like time limits, awards in form of points and achievements, difficulty levels, leader boards or randomness of input data [14].

The quality of target game can be described by two parameters: average lifetime play (the time that average player spent playing the game) and throughput (average number of problems solved per hour of playtime) [4]. Simko [19] also pointed out that GWAP should be characterized also by the total number of players that took part in the game.

III. Wikipedia - WordNet Connections Validation

During the Colabmap project we created a set of mappings between English Wikipedia articles and WordNet synsets. Sample mappings are presented in Table I. Each mapping consists of a WordNet synset, definition of the synset and the title of Wikipedia article with special characters coded using RFC 3986. Such mappings, when proved to be correct, will allow formalization of Wikipedia structure. The obtained set of mappings contained algorithmically created 54475 connections that required verification. Tempted by the results obtained during the Samsung's survey we decided on implementing a GWAP for validation of those connections.

The originally obtained mappings were extended with three additional "next best" mappings with the idea of presenting the user a question (definition of a synset) with 4 possible answers (Wikipedia article titles). At the beginning the 3 other answers were randomly selected from the set of Wikipedia's pages but such approach quickly proved to be incorrect as the "next best" mappings were not related at all to the question. Instead we used Wikipedia search functionality to select alternative answers (according to Wikipedia) following the Algorithm 1.

Algorithm 1 Algorithm for selecting alternative answers

```
for all synonyms of WordNet synset do
    Read the synonym
    Perform a Wikipedia search using the synonym
    Store 3 top elements from search results
end for
```

The example of extended mappings are presented in Table II. For the tests we randomly selected 100 synsets from our database to limit the time needed to gather the results and verify the viability of the game.

IV. TGame

We decided to implement TGame[3] ("Tagger Game") as a 2D platform game following the output-agreement model.

[2]http://en.wikipedia.org/wiki/Google_Image_Labeler

[3]https://play.google.com/store/apps/details?id=pl.gda.eti.kask.tgame, http://kask.eti.pg.gda.pl/tgame/

TABLE I
SAMPLE WORDNET – WIKIPEDIA MAPPINGS

Synset (WordNet)	Definition (WordNet)	Article (Wikipedia)
Sept. 11, September 11, 9-11, 9/11, Sep 11	the day in 2001 when Arab suicide bombers hijacked United States airliners and used them as bombs	September_11
interval, time interval	a definite length of time marked off by two instants	Time
ice age, glacial epoch, glacial period	any period of time during which glaciers covered a large part of the earth's surface	Ice_age Glacial_period
man hour, person hour	a time unit used in industry for measuring work	Man-hour
entity	that which is perceived or known or inferred to have its own distinct existence (living or nonliving)	Entity
French leave	an abrupt and unannounced departure (without saying farewell)	French_leave
hunt, hunting	the pursuit and killing or capture of wild animals regarded as a sport	Huntingdon
blindman's bluff, blindman's buff	a children's game in which a blindfolded player tries to catch and identify other players	Blind_man%27s_bluff
landler	a moderately slow Austrian country dance in triple time	L%C3%A4ndler
coup d'oeil, glimpse, glance	a quick look	Coup_d%27%C5%93il

TABLE II
SAMPLE OF EXTENDED MAPPINGS, THE ORIGINAL MAPPING IS EMPHASIZED

Synset (WordNet)	Articles (Wikipedia)
Sept. 11, September 11, 9-11, 9/11, Sep 11	*September 11*, 9/11 Commission, 9/11 conspiracy theories, United Airlines Flight_93
interval, time interval	*Time*, Interval (music), Interval, Interval (mathematics)
ice age, glacial epoch, glacial period	*Ice age*, Pleistocene, Wisconsin glaciation, Gravettian
man hour, person hour	*Man-hour*, Hourman, Man of the Hour, 24 Hours of Le Mans
entity	*Entity*, Administrative divisions of Mexico, Administrative division
French leave	*French leave*, French leave (disambiguation), Desertion, French Leave (1930 film)
hunt, hunting	*Huntingdon*, Hunting, Fox hunting, Seal hunting

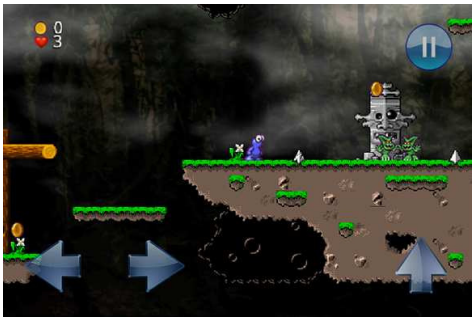

Fig. 1. TGame.

Fig. 2. Checkpoint activation.

We chose Android platform as a test environment due to its popularity and ease of access for users and developers. The game implements standard features like different levels and collectibles (coins, hearts), need of finishing one level before the other one is accessible. The player is encouraged to replay levels by a simple point system that awards the player for killing monster, gathering stars and hearts (Figure 1).

A. Tying questions with the game

One of the biggest challenge is to properly include the mappings (or any type of a general question) into the game. We tried to implement the questions to be as non intrusive as possible but still easy to stumble upon. In TGame the verification of mappings is done when the player wants to activate a checkpoint (a respawn place when player is moved when killed). To activate the checkpoint player needs to answer the question provided by marking the correct mapping (Figure 2). When the answer is identical to the one stored in the database the checkpoint is activated. If the player chose other answer then the checkpoint is not enabled. When the player is certain that he or she selected a correct answer then he or she can report his or her answer using the proper option in the pause menu. The checkpoint is then activated for one use.

From the technical point of view the communication between the client and the server goes as follows. Each client upon first connection downloads pack of configurable number of questions and possible answers so connection to the Internet

is required only at first start of application and later at user chosen intervals. Whenever possible the game sends gathered results with statistical information and downloads new pack of questions (if needed).

B. Answer Verification

The process of reporting wrong answers requires explicit action from the user. It was designed to require some activity but not too much so not to discourage the users. Very easy access to submission would encourage people to skip reading the question and just submitting information about wrong answer. In general the game has to be paused and proper menu have to be selected. Only the last question can be reported.

Furthermore when submitting results also time elapsed between displaying the question and selecting the answer is also submitted. Such extensions allows us to eliminate submissions that i.e. are so short that the user would not be able to read the question. Randomly selected batch of questions required on average 5 seconds to be properly read and understood by players. We decided to discard all answers that took less than 4 seconds (8% of all results).

The answers that the players gave (correct, incorrect and reported) are later compared with the one calculated by Colabmap algorithms. All the selected answers are visible in administrators panel of the server side of the solution and can be exported using csv format to an external tool.

C. Results Analysis

During first two months of tests the game was downloaded by 25 people, mainly students and friends (currently according to Google Play web page there are between 50 and 100 downloads without any additional advertising). The original 25 players gave 626 answers for 100 questions. The game run 10 hours in total on multiple devices. Each player solved 44,42 questions per hour. At this rate, with average playtime of each player at 5 hours, we would need minimum 2500 players to answer each question at least once. Judging by other similar games available on Google Play such number can be achieved with proper advertising of the product given the user base and popularity of mobile games.

During the evaluation of the proposed solution we faced two main type of problems. In some cases the additional answers generated using the Wikipedia search functionality provided very similar pages which introduced difficulty in choosing the correct one. Selection of 100 random questions also introduced problem with high variety of difficulty level among questions. It became obvious that some of them require expert-level knowledge. Examples of questions belonging to those two groups are presented in Table III. The column "Answer 1" contains the correct mapping. Furthermore the type of game implemented (a simple platform game) did not match the questions asked. In future work we will reorganize the questions to Yes/No/Unsure form to lower the difficulty level and implement other type of games to better match the type of mappings that are verified.

TABLE III
SAMPLE QUESTIONS WITH HIGH LEVEL OF DIFFICULTY

The Question	Answer 1	Answer 2	Answer 3	Answer 4
Asiatic nut trees: wing nuts	Pterocarya	Pterocarya fraxinifo-lia	Pterocarya stenoptera	Cyclocarya
a colorless flammable volatile liquid hydrocarbon used as a solvent	Xylene	O-Xylene	P-Xylene	Xylene cyanol
a former large county in northern England	Yorkshire	Yorkshire 6	Yorkshire captaincy affair of 1927	South Yorkshire Fire and Rescue
fine porcelain that contains bone ash	Bone china	Aynsley China	Bisque (pottery)	Porcelain

V. MAPPING UPDATE

We tried three approaches for deciding whether the mapping, based on the answers provided by the players, should be updated or not:

- The mapping was considered correct when 75% of the player answers agreed. This approach however did not give any results as only 50% of original mappings managed to get enough answers, none of the incorrect mappings were marked as correct.
- The mapping was considered correct when at least 50% of player answers agreed. In this case 64% of all mappings were marked as correct which covered 75% of all mappings marked as correct in our database. Unfortunately this method generated some false positives.
- The mapping which gathered the most of the player answers was considered correct. In this case 74% of all mappings were marked as correct witch covered 80% of all mappings originally marked as correct in our database. This method also generates false positives.

Currently in our solution we implemented the third method as it provided the best results. Still this method does not allow us to automatically state whether the given mapping is correct or not. However "problematic" mappings are clearly pointed out by the players (by either majority of wrong answers or reports). Such mappings can than be verified manually by experts. In our further work we plan on extending the procedure with additional parameters like user reputation, level and field of education, history of answers etc. which should improve the level of trust that can be put in user the generated answers.

During the evaluation period the players submitted 17 mappings update requests regarding 12 questions. Sample reports are presented in Table IV. The corrected mappings are emphasized.

TABLE IV
UPDATED MAPPINGS

WordNet Definition	Original mappings	Other available answers
an advanced student or graduate in medicine gaining supervised practical experience ('houseman' is a British term)	Internet Movie Database	Houseman, Julius Houseman, *Internship (medicine)*
large voracious aquatic reptile having a long snout with massive jaws and sharp teeth and a body covered with bony plates	Crocodile tears	*Crocodile*, Crocodylus, Schnappi
(elections) more than half of the votes	Supermajority	*Majority*, Simple majority, Absolute majority

VI. CONCLUSIONS AND FURTHER WORKS

We proposed a platform for verification of the results of heuristic algorithms. Currently verification of mappings is supported. We verified the solution using Wikipedia – WordNet mappings and managed to get some promising results and were able to correct some of the mappings. The problems that still need to be solved include better formulation of the question and the trust that the system can put in answers provided by the users.

We also plan on extending the proposed solution by generalizing it for other type of tasks, inclusion of different clients, not only game based, designed for certain types of questions or with required user knowledge in mind. We are also currently implementing social features like achievements and leader boards that should lure more players and create a wider user base. In case of Wikipedia - WordNet mappings we plan on tagging questions with difficulty levels and include them in a quiz-like game similar to "Fifteen to One"[4] or "1 of 10"[5]). Such type of client could be more suitable for such defined problems. The TGame can be a great application for crowd base image tagging or a client when the questions will be redesigned to a Yes/No format.

Our research shows that popularity of computer games and wide availability of devices that can be used for playing at any time makes GWAPs an approach that has some unexplored potential. Our first implementation, despite its drawback, shows that this potential is relatively easy to unlock. Even for small user base we managed to find some errors in our mappings. Implementation of different client applications more fitting the types of tasks needed to be done (image tagging, sound analysis etc.) and careful question formulation should enable us to fully unlock the possibility of crowdsourcing based task execution. When succeeded such possibility can be of great help to researchers around the world as it reduces resources and time needed to verify the results of designed algorithms and implementations. Furthermore it can be implemented as an alternative to in app purchases or advertisements. This way the users can be provided with content with their work be treated as another means to "pay" for it.

[4]http://en.wikipedia.org/wiki/Fifteen_to_One

[5]http://pl.wikipedia.org/wiki/Jeden_z_dziesi%C4%99ciu

REFERENCES

[1] *Prawie połowa Polaków gra codziennie w gry wideo (in Polish)*, URL http://samsungmedia.pl/pr/223805/prawie-polowa-polakow-gra-codziennie-w-gry-wideo. [Online], Biuro Prasowe Samsung Electronics Polska Sp. z o.o.

[2] U. Świerczyńska-Kaczor and J. Wachowicz, "Student response to educational games – an empirical study," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha, L. Maciaszek, Ed. IEEE, 2013, pp. pages 1293–1299.

[3] D. P. Anderson and G. Fedak, "The computational and storage potential of volunteer computing," in *Cluster Computing and the Grid, 2006. CCGRID 06. Sixth IEEE International Symposium on*, vol. 1. IEEE, 2006, pp. 73–80.

[4] L. Von Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92–94, 2006.

[5] J. Howe, "Crowdsourcing: A definition," *URL http://www.crowdsourcing.com/cs/2006/06/crowdsourcing_a.html.[Online]*, p. 29, 2006.

[6] V. Osinska, A. Jozwik, and G. Osinski, "Mapping evaluation for semantic browsing," in *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 5. IEEE, 2015, pp. 329–335. [Online]. Available: http://dx.doi.org/10.15439/2015F50

[7] D. Wightman, "Crowdsourcing human-based computation," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*. ACM, 2010, pp. 551–560.

[8] A. Kosorukoff, "Human based genetic algorithm," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, vol. 5. IEEE, 2001, pp. 3464–3469.

[9] J. Simko and M. Bieliková, "Games with a purpose: User generated valid metadata for personal archives," in *Semantic Media Adaptation and Personalization (SMAP), 2011 Sixth International Workshop on*. IEEE, 2011, pp. 45–50.

[10] R. Korytkowski and J. Szymanski, "Collaborative approach to WordNet and Wikipedia integration," in *The Second International Conference on Advanced Collaborative Networks, Systems and Applications, COLLA*, 2012, pp. 23–28.

[11] J. Szymański, "Mining relations between Wikipedia categories," in *Networked Digital Technologies*. Springer, 2010, pp. 248–255.

[12] ——, "Words context analysis for improvement of information retrieval," in *Computational Collective Intelligence. Technologies and Applications*. Springer, 2012, pp. 318–325.

[13] J. Szymański and W. Duch, "Self organizing maps for visualization of categories," in *Neural Information Processing*. Springer, 2012, pp. 160–167.

[14] L. Von Ahn and L. Dabbish, "Designing games with a purpose," *Communications of the ACM*, vol. 51, no. 8, pp. 58–67, 2008.

[15] ——, "Labeling images with a computer game," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2004, pp. 319–326.

[16] L. Von Ahn, S. Ginosar, M. Kedia, and M. Blum, "Improving image search with phetch," in *Acoustics, speech and signal processing, 2007. icassp 2007. ieee international conference on*, vol. 4. IEEE, 2007, pp. IV–1209.

[17] L. Von Ahn, R. Liu, and M. Blum, "Peekaboom: a game for locating objects in images," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 55–64.

[18] E. L. Law, L. Von Ahn, R. B. Dannenberg, and M. Crawford, "TagATune: A game for music and sound annotation," in *ISMIR*, vol. 3, 2007, p. 2.

[19] J. Simko, "Semantics discovery via human computation games," *Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications*, p. 286, 2013.