HappyLabel: A Human Computation Game with Predefined Levels of Categorization

Yinan Na, Zheng Shen
Stanford University
450 Serra Mall
Stanford, CA 94305, USA

ABSTRACT

We present HappyLabel, a human computation game incorporates predefined levels of categorization. HappyLabel's approach of incentivizing players to label image progressively produces more informative and structured meta tags of images. Online experiments demonstrate that our design improves the efficiency of image labeling, enables deep labeling with specific information, and encourages people to play.

ACM Classification: H5.2 User Interfaces. – Human Computation.

Keywords: ESP game, Human in the loop, Human computation, Cognitive category

INTRODUCTION

The ESP game[1], initially introduced by von Ahn, is designed to incorporate crowd intelligence to work on image labeling tasks, which is still difficult for automatic computation-based system. And its success led a series of inventions of similar applications as Human computation games, or Games With A Purpose (GWAP)[1,2,3,4].

Despite of influential success, ESP game bears shortcoming in design that encourages players to label on "obvious" words[1,5,6,7] to the image, and thus information provided alongside is less informative. Many works have been done to address this problem[5,6,7]. Most of them focus on human incentives as well as language model in either theoretical analysis or statistical modeling.

In this paper, we propose a different and complementary approach to work on this problem. Rather than working human-side in game playing, we incorporate predefined category structure to create an interactive hybrid human-computer computation game inspired by Human-in-the-loop (HITL) [8], and apply dynamic score systems, to encourage participants to progressively provide informative answers. Three hypotheses are aligned with such mechanism design: (i) Using predefined category structure, participants' guess can be regulated, and thus with higher efficiency; (ii) Progressive options enable participants to provide labels in more specific categories, and thus more informative; (iii) Such design can also make the game fun to play. To test these hypotheses, we introduce the game HappyLabel to implement our design via invited testing and facebook promotion.

CS376 '13, June 10–13, 2013, Stanford, California, Unite States.

RELATED WORK

Study And Improvement Of ESP Game

GWAP proposed to use game design, exploit people's intrinsic motivators, involve people in task voluntarily, and harvest contributions as side-products of their playing. To ensure the quality of outcomes, several properties are applied: (i) players share the common goal of "agreeing" on certain things, (ii) players are matched randomly, (iii) no communication is allowed[3], thus the best strategy for players is to provide truthful answer.

Original design of ESP game is not resistant to people's tendency to guess easy and generic words. Rigorous theoretical study also proved this shortcoming[7]. von Ahn introduced taboo words to alleviate this problem[1]. However taboo words are static, and may present additional information that bias guessing[6]. To improve the taboo,

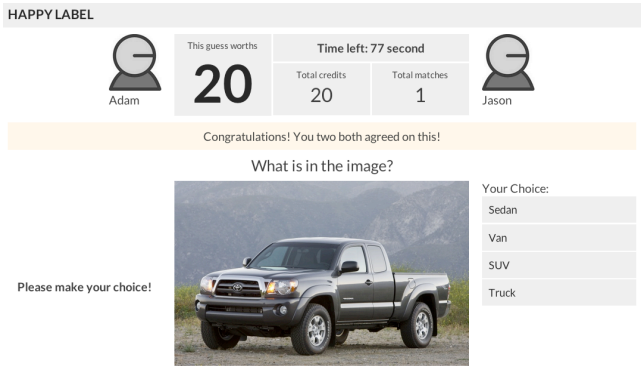

Figure 1: The interface of game play of HappyLabel

KissKissBan[6] involved a third person named blocker to generate taboo accordingly, [5] proposed to using hidden taboo to avoid biased information. Their approach may encourage more diverse labels, however still cannot guarantee specificity of information.

Another approach is using dynamic incentive stimulus to encourage guessing more informative words. Google introduced different scores depending on "specificity"of words [5,7]. Later, Webel[5] proposed proportional scoring system can work better with their language model that predicts users' labels. These works did not regulate the cognitive space for image guessing, people still bear challenges to search answer from very broad knowledge background.

Human In The Loop

Human-in-the-loop (HITL) defined as a model that requires human interaction. It enables human to change the outcome of a process, and thus can leverage human's contribution to drive up performance in human-computer framework. ESP game is one example, since its outcomes are good materials for training machine to label images. There also exists many other processing work in computer science with human inputs[8,9,10,11]. In Branson's work[8], an interactive, hybrid human-computer method for image classification is introduced.

Cognitive Categories

One way of objective naming is by categorization, and is organized into taxonomy, which are in turn organized into levels of categorization[12] Class manifest themselves in three levels of categorization based on inclusion and specificity: namely the superordinate level, the basic level and the subordinate level, and In most cases, the difficulty for human to identify these levels is progressive from the former to the latter.

Our approach is inspired by work of HITL[8], and introduce cognitive categories as well as dynamic scoring systems into current ESP game model.

HAPPYLABEL

Happy Label has similar player pairing and interacting mechanism as ESP games: (i) players are randomly paired, (ii) they cannot communicate, and (ii) they earn scores via reaching consensus. However, when guess about one image, instead of typing in words, players are provided with at most 4 rounds of multiple choices with increasing difficulty. Players guess the object in image via selection. When player correctly answering successive multiple-choice selections, an image can be attached with a more specific, informative labels. Main features to facilitate such game mechanism including: Data tree structures, dynamic score systems, and shuffled multi-choice lists.

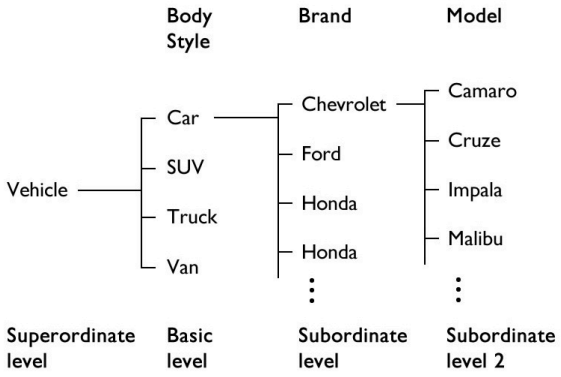

Figure 2: The tree structure of category levels

Category Tree Structures

We use a tree structure to store the predefined category in levels. This category tree is later used to generate options automatically for each labeling task. Take vehicle labeling as instance. The root node of tree has the label "Vehicle" as the superordinate category level. Under "Vehicle" node, there are child nodes as body style, including "Sedan", "SUV", "Truck" and "Van". These are of basic category level. Child nodes under "Car" are manufacture brands, like "Chevrolet" and they are of first subordinate level category. Under brands , there are child nodes as specific models, like "Camaro" in second subordinate level category. This category tree has four levels (or with depth 4).

In our game, each image will be labeled starting from the root node with options of "Sedan", "SUV", "Truck" and "Van". If players reach a match on their selection, they will go to the next level and system generates new options from child nodes of the new root (the matched node) correspondingly. This labeling process continues until there is a mismatch or it hits the bottom level of the tree where the most specified options reside.

Scores as Incentives

Difficulty of player recognizing is increasing from basic-level category to sub-level categories. To encourage people to challenge such difficulties, a dynamic score system is utilized. Players are rewarded on consensus they reached in selection, and scores are increasing according to the level of category. The reward of matching in the subordinate level is twice to the parent level, and since players have to keep matching in parent levels before reaching the subordinate level, the accumulated score of one image is even more remarkable. Such reward mechanism is similar to the game in TV show *Who Wants To Be Millionaire*[13].

Dealing with Cheating and Abuse

Designing crowdsourcing algorithms is often like designing a user interface that will keep a user "in bounds" on your application. To prevent players from cheating and abuse,

we carefully choose the choices with the requirements that: options in the same level should have same level of inclusion and descriptiveness, no option like "others" or "I don't know" is offered. This ensures players have no relative easy options in a round of multi-choice list, and thus cannot collude. In addition, options in a round are shuffled when providing to players, so that 2 players cannot find pattern of partner's behavior via the order of options.

IMPLEMENTATION

We implemented HappyLabel as a browser game. The client side was implemented using HTML5, CSS3 and Javascript. The backend was built using Firebase. We stored all the data (including game information, player information, leaderboard information, image labels, category tree) in Firebase. Any changes (read or write) made to that data are automatically synchronized with the Firebase cloud and with all clients within milliseconds.

FIELD DEPLOYMENT AND EVALUATION

To evaluate the efficiency and accuracy of the image labeling with our design, we published HappyLabel online to collect data. And in addition, we conducted a user interview involving 5 players to evaluate the gameplay of HappyLabel.

The evaluation is base on the 130 images of vehicles. Category tree structure is defined based on categories from vehicle vaulation website Edmunds.com[...], pictures of vehicles associate to the category tree are also selected from the same domain. Each game lasts 1 minutes and half. To ensure people to pair up and play synchronously, we designed the feature that one player generates link and invite friends to join

Efficiency

The test period lasts 3 days, and during this time, 42 games were played, 105 images were labeled in total 650 rounds of guessing. On average, there 15.44 labels (std. 5.93) were created in each game, including 11.31 matched (std. 6.10), 4.12 unmatched (std. 3.37). The results shows the production of labels is much higher than ESP games. The average time to match specific words is also faster than guessing on generic and easy words in ESP game. One reason is that the options are highly regulated, thus players need much less time to decide input and reach a match.

Informative Labeling

For each image, on average 1.476 times it was guessed. In the images that were labeled. all images reach basic-level labels, and only 11 images were labelled wrong at least once, only 2 totally wrong. The correctness is 89%. In the first subordinate level, 88 images were labeled, 83% of all labeled images, and the correctness is 96.5%, In the second subordinate level. 41 images were labeled, 39% of all labeled images, 51.2% are correct.

Though the average time one image was guessed is low, considering that players are randomly selected, the result can still be considered as significant. This result shows that with the predefined category structures, images can be labeled into more specific categories with high accuracy.

Gameplay Feedback

5 Players (2 females, 3 male) that participated in interviews are aged from 20 to 30, including 4 students, 1 employees in technology company. All of them played at least 2 times of the game. For features that make the game exciting, 3 people thought progressive scoring system, all people thought time constraint. For difficulty due to the domain knowledge (knowledge about vehicles), 3 people acclaimed that they do not have domain knowledge, and options in second subordinate level were really challenging, Last but not least, 4 of them would like to play more.

DISCUSSION & FUTURE WORK

In general, The results of our evaluation support well the three hypotheses we created at the beginning of the paper. There are still several interesting parts worth mention. The accuracy in first subordinate level options is even higher than the basic level options. This might because people's knowledge is not always align with the taxonomy. Another similar case is not many people know the species the dog belongs to, even though such species is actually more general in inclusion and descriptiveness. Thus when design similar games, the selection of category levels needs more consideration of people's common knowledge. Similarly, other characteristics of category levels are also worth further study: including the depth of levels, and whether to include options for user to input own tags.

The game mechanism of HappyLabel requires predefined category structure. and thus creating robot system for such interactive option generating to enable asynchronous game playing is more challenging than that of original ESP game. How to implement current game with Mechanical Turk is also worth trying to get more data for further analysis.

In reality, Our game design has a wide range of application. For instances, online shopping websites often need to categorize items into deep taxonomy trees, our design can works as an attractive feature to not just label items, but also engage visitors with fun interactivity. Because such game requires no text input, it is very easy for people to play on mobile devices, and harvest much more labor from interfaces on which people spend more leisure time than they do on computers. Furthermore, it can also works as post process of current ESP games to label images roughly

classified with more informative tags from deep cognitive categories.

CONCLUSION

This study introduces HappyLabel, a human computation game implements predefined levels of categorization for labelling images with more specific words from progressive cognitive category levels. In the game implementation and testing, We demonstrated that such design improves the efficiency of labeling with regulated user guessing, enables deep level labeling through interactive progression in label options, and ensures enjoyable gameplay through dynamic scoring rewards. By incorporating predefined taxonomy and interactive system feedback, such game design can work as complementary of existing ESP games or other image labeling process.

ACKNOWLEDGMENTS

We really appreciate Michael, and TA's help during the whole project. We thank people who participate in our experiments and provide precious advices during the game design.

REFERENCES

1. von Ahn, L. & Dabbish, L., 2004. Labeling images with a computer game. In Proc. CHI '04: 319–326.
2. von Ahn, L., 2006. Games with a purpose. IEEE Computer Magazine, 39(6):92–94.
3. von Ahn, L. & Dabbish, L., 2008. Designing games with a purpose. Communications of the ACM 51, 58–67.
4. von Ahn, L., Ginosar, S., Kedia, M., Liu, R., Blum, M., 2006a. Improving accessibility of the web with a computer game, in: Proceedings of the 2006 Conference on Human Factors in Computing Systems (CHI), pp. 79–82.
5. Weber, I., Robertson, S., and Vojnovic, M., 2009. Rethinking the ESP game. in extended abstracts CHI '09
6. Ho, C.J., Chang, T.H., Lee, J.C., jen Hsu, J.Y., Chen, K.T., 2009. Kisskissban: A competitive human computation game for image annotation, in KDD-HCOMP.
7. Jain, S., & Parkes, D. C. 2008. A game-theoretic analysis of games with a purpose. In Proc. 4th Intl. World Internet and Network Economics (WINE)
8. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., & Belongie, S. 2010. Visual recognition with humans in the loop. In Computer Vision–ECCV 2010, pp. 438-451.
9. Zhou, X., Huang, T., 2003. Relevance feedback in image retrieval: A comprehensive review. Multimedia Systems 8, 536–544
10. Tong, S., Koller, D., 2002. Support vector machine active learning with applications to text classification. JMLR 2, 45–66
11. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T., 2007. Active learning with gaussian processes for object categorization. In: ICCV, pp. 1–8
12. http://cogling.wikia.com/wiki/Levels_of_categorization
13. http://millionaire.itv.com/home/