An Analytical Study of Puzzle Selection Strategies for the ESP Game*

Ling-Jyh Chen[1], Bo-Chun Wang[1], Kuan-Ta Chen[1], Irwin King[2], and Jimmy Lee[2]

[1] Institute of Information Science, Academia Sinica
[2] Department of Computer Science & Engineering, Chinese University of Hong Kong

Abstract

"Human Computation" represents a new paradigm of applications that take advantage of people's desire to be entertained and produce useful metadata as a by-product. By creating games with a purpose, human computation has shown promise in solving a variety of problems that computer computation cannot currently resolve completely. Using the ESP game as an example, we propose a metric, called system gain, for evaluating the performance of human computation systems, and also use analysis to study the properties of the ESP game. We argue that human computation systems should be played with a strategy. To this end, we implement an Optimal Puzzle Selection Strategy (OPSA) based on our analysis to improve human computation. Using a comprehensive set of simulations, we demonstrate that the proposed OPSA approach can effectively improve the system gain of the ESP game, as long as the number of puzzles in the system is sufficiently large.

1. Introduction

"Human Computation" represents a new paradigm of applications that take advantage of people's desire to be entertained by outsourcing certain steps of the computational process to humans [2–4]. In [5], Ahn proposed the use of human computation to create *games with a purpose* that provide entertainment and produce useful metadata as a by-product. By exploiting "human cycles" in computation, human computation has shown promise in solving a variety of problems, such as image annotation and commonsense reasoning, which computer computation has been unable to resolve completely thus far.

In this work, using the ESP game as an example, we define a metric, called *system gain*, to evaluate the performance of human computation systems. The proposed metric considers two factors: the number of puzzles that have been played in the system, and the average outcomes produced by each puzzle. Both factors are critical for human computation systems, but unfortunately they do not complement each other. We believe that human computation systems should be *played with a strategy*. Specifically, based on our analysis, we propose an *Optimal Puzzle Selection Algorithm* (OPSA) that can maximize the system gain by properly accommodating the two opposing factors. Using a set of simulations, we investigate the properties of the ESP game, and evaluate the proposed OPSA scheme on two widely used schemes, namely the *Random Puzzle Selection Algorithm* (RPSA) and the *Fresh-first Puzzle Selection Algorithm* (FPSA). The results demonstrate that, with the OPSA scheme, the ESP system yields a much better system gain than the two compared schemes.

The remainder of this paper is organized as follows. Section 2 contains a review of related works on human computation systems and describes the rules of the ESP game. In Section 3, we present our analysis of the ESP game. In Section 4, we compare three puzzle selection algorithms for the ESP game, namely the RPSA, FPSA, and OPSA schemes. Section 5 presents a comprehensive set of simulation results, which we analyze and explain in detail. We then summarize our conclusions in Section 6.

2. Background

"Human Computation" was pioneered by Luis von Ahn and his colleagues, who created games with a purpose [5] that people play voluntarily and produce useful metadata as a by-product. By taking advantage of people's desire to be entertained, Human Computation has shown promise in solving some problems that computer computation cannot currently resolve completely. The online ESP Game [6] was the first human computation system, and it was subsequently adopted as the Google Image Labeler [1]. When a user logs into the system, he/she is automatically matched with a random partner. The two players do not know each other's identity as they cannot communicate.

*This research was supported in part by Taiwan E-learning and Digital Archives Programs (TELDAP) sponsored by the National Science Council of Taiwan under NSC Grants: NSC 96-3113-H-001-012.

Initially, a randomly selected image is displayed to both players simultaneously. The players then input possible words to label the image until an "agreement" is reached (i.e., the same word is entered by both players), and a bonus score is awarded to each player based on the '*quality*' of the agreed word. In practice, the '*quality*' of a word is measured by its popularity; generally, words that are more popular receive lower scores. After the players agree on a word, they are shown another image. In each game, they have two and a half minutes to label 15 images. The word on which the two players agree becomes the label of the image, and it can not be used the next time that image is displayed in another game (the word is called a "taboo" word of the image). The rationale for using taboo words is to ensure that each image is labeled with a variety of words.

To be effective, the ESP game tries to collect outcomes with the largest possible aggregated score for each puzzle (image), and thus needs as many distinct puzzles as possible to be played. There is a trade-off between these two aspects. On the one hand, the system prefers to take as many labels as possible for each puzzle, which will result in the playing of fewer distinct puzzles; on the other hand, the system prefers that each puzzle is played only once, which can lead to the playing of the maximum number of puzzles. Thus, an optimal puzzle selection strategy that can accommodate the two goals is highly desirable. To this end, we propose a metric to evaluate the system gain of the ESP game, and analyze the puzzle selection problem.

3. The Analysis of the ESP Game

Let N be the number of the puzzles that have been played at least once in the system, and let S be the total score of all the agreed labels. We define the system gain, G, of the ESP game as follows:

$$G = ln(N) \times ln(S/N). \tag{1}$$

Clearly, G increases as the number of the games played increases, and/or as the average total score (per puzzle) increases. Suppose that, in the system, each puzzle has the potential to yield K labels in total, and each tag is associated with one positive score value based on its popularity. For simplicity, we assume there are totally X distinct scores (i.e., $S_1, S_2, S_3, ..., S_X$) in the system, and $S_i = e^i$. Moreover, we assume that K_i labels have the score S_i, and $K_i = e^{X-i}$. Therefore the total number of potential labels per puzzle (K) can be derived by Eq. 2, and the expected score of each tag ($E[S]$) can be obtained by Eq. 3.

$$K = \sum_{i=1}^{X} K_i = \frac{e^X - 1}{e - 1} \tag{2}$$

$$E[S] = \frac{\sum_{i=1}^{X} e^i e^{X-i}}{K} = \frac{e^X (e-1) X}{e^X - 1} \tag{3}$$

Suppose the N puzzles have been played T rounds in total, and each puzzle has been played r times on average ($r = T/N$). We can then rewrite Eq. 1 as follows:

$$\begin{aligned} G &= ln(T/r) \times ln(E[S] \times r) \\ &= -\left(ln(r) - \frac{ln(T) - ln(E[S])}{2}\right)^2 + C, \end{aligned} \tag{4}$$

where C is a constant and equal to $ln(T)ln(E[S]) + \left(\frac{ln(T)-ln(E[S])}{2}\right)^2$. Note that C also represents the largest possible system gain that occurs when $r = e^{\frac{ln(T)-ln(E[S])}{2}}$.

4. Puzzle Selection Algorithms

In this section, we compare three puzzle selection algorithms for the ESP game, namely the *Random Puzzle Selection Algorithm* (RPSA), the *Fresh-first Puzzle Selection Algorithm* (FPSA), and the proposed *Optimal Puzzle Selection Algorithm* (OPSA). We use P to denote the set of all puzzles in the system, and define the following three functions used by the puzzle selection algorithms: 1) $Select\_Random(P)$, which randomly selects a puzzle from the input puzzle set P; 2) $Select\_Played(P)$, which selects the puzzle in the input puzzle set P that has been played most frequently; and 3) $Select\_Fresh(P)$, which selects the puzzle in the input puzzle set P that has been played least frequently. We present the three algorithms in the following.

4.1. RPSA and FPSA

The *Random Puzzle Selection Algorithm* (RPSA) selects a puzzle at random from the puzzle pool P, using the function $Select\_Random(P)$, in each round, and it provides the baseline performance of the ESP game in this study. On the other hand, the *Fresh-first Puzzle Selection Algorithm* (FPSA) selects the puzzle that has been played least frequently, using the function $Select\_Fresh(P)$. It is a greedy, heuristics-based approach that tries to maximize the first component of Eq. 1.

4.2. The Proposed Scheme: OPSA

In the proposed *Optimal Puzzle Selection Algorithm* (OPSA) for ESP games, N denotes the number of puzzles that have been played in the system, E denotes the expected score of each label, and T is the total number of rounds that have been played. In addition, r denotes the optimal number of rounds; and for each entry p of P, $p.r$ represents

Algorithm 1 The Optimal Puzzle Selection Algorithm.

Function OPSA
$T \Leftarrow T + 1$; $r' \Leftarrow \lceil e^{\frac{ln(T)-ln(E)}{2}} \rceil$
if $r' > r$ **then**
  for each p in P_2 **do**
    if $p.r < r'$ **then**
      Move p from P_2 to P_1
    end if
  end for
  $P_1 \Leftarrow P_1 \bigcup P_2$; $r \Leftarrow r'$
end if
if $\{P_1\}$ is NOT empty **then**
  $p \Leftarrow Select\_Played(P_1)$; $p.r \Leftarrow p.r + 1$
  if $p.r = r$ **then**
    Move p from P_1 to P_2
  end if
else
  if $\{P_0\}$ is NOT empty **then**
    $p \Leftarrow Select\_Random(P_0)$; $p.r \Leftarrow 1$
    if $p.r < r$ **then**
      Move p from P_0 to P_1
    else
      Move p from P_0 to P_2
    end if
  else
    $p \Leftarrow Select\_Fresh(P_2)$; $p.r \Leftarrow 1$
  end if
end if
Return p

the round number in which the puzzle p was played. Suppose the puzzle set P_0 contains all the puzzles that have not been played; P_1 contains all the puzzles that have been played at least once, but less than r rounds; and the set $P_2 = P - P_0 - P_1$ contains the other puzzles. We detail the OPSA algorithm in Algorithm 1.

5. Evaluation

5.1. The Optimal r

In the first set of simulations, we evaluated the accuracy of our analytical model in determining the optimal r value for the ESP game. We assumed that the number of puzzles in the system was infinite, and all of them were unsolved at the beginning of the simulation (i.e., no labels were discovered for any puzzles). Moreover, we set the total number of game rounds played (T) at 10,000. Figure 1 shows the evaluation results in terms of system gain for r values between 2 and 230, when the maximum score value X was fixed at 6. In the figure, the analysis curve is derived by Eq. 4, where $E[S]$ can be obtained by Eq. 3. It is equal to 10.3353 when $X = 6$. We observe that the analysis curve matches the simulation curve very well, and the optimal r values (i.e., those that yielded the largest system gain) of the two curves are also comparable.

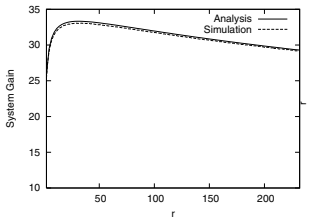

Figure 1. Comparison of the system gain under various r settings in both the simulations and the analysis. ($X = 6$ and $T = 10,000$)

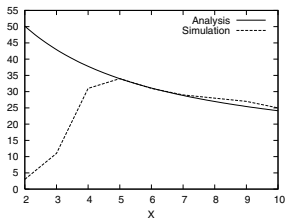

Figure 2. Comparison of the optimal r values derived by simulations and analysis, where $T = 10,000$ and X varies between 2 and 10.

Additionally, we varied the maximum score value X in the range 2 to 10 and compared the derived optimal r values using both simulations and analysis, as shown in Figure 2. The results indicate that the analysis curve only matches the simulation results well when $X \geq 5$. The reason is that, in the analytical model, the optimal r value is larger than the total number of potential tags per puzzle (K in Eq. 2) when $X < 5$. Thus, this model can not be used when $X < 5$ because the optimal number of rounds per puzzle is larger than the number of tags that a puzzle has in the system.

From Figure 2, we also observe that, when $X > 5$, the optimal r value decreases as the X value increases. This confirms our intuition that the value of $E[S]$ increases as X increases. As a result, the optimal r value will decrease as X increases. We find that if there are several different scores in the system, more rounds of each puzzle must be played in order to achieve a better overall system gain.

5.2. The Relationship between T, N, and r

Next, we evaluate the relationship between the total number of game rounds T, the number of played puzzles N, and the number of game rounds required to maximize the system gain r in the proposed Optimal Puzzle Selection Algorithm. Figures 3 and 4 show the comparison results of r and N with various T values in the range 200 to 20,000 (X is fixed at 6).

Figures 3 and 4 show that our analytical model matches the simulation results very well in all cases. In addition, we observe that both of the r and N values increase as the value of T increases. There are two reasons for this phenomenon: a) as the total number of game rounds increases, each puzzle tends to take more labels from the system; and b) a larger number of puzzles are played. Since $N = T/r$,

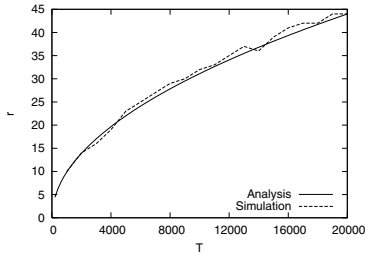

Figure 3. Comparison of the optimal r values derived by simulations and analysis, where $X = 6$ and T varies between 200 and 20,000.

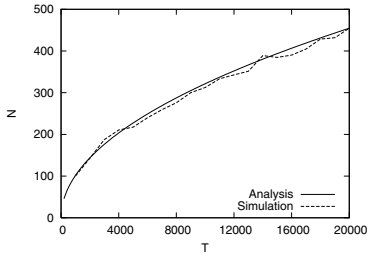

Figure 4. Comparison of the N values derived by the simulations and analysis, where $X = 6$ and T varies between 200 and 20,000.

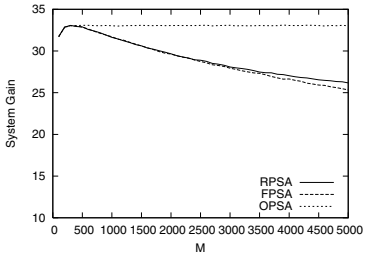

Figure 5. Comparison of the system gain achieved with various numbers of puzzles, where T is fixed at 10,000 and X is set to 6.

the results confirm that the proposed OPSA approach can effectively balance the two goals, i.e., maximize the number of games played, while identifying as many labels per puzzle as possible.

5.3. Comparison of RPSA, FPSA, and OPSA

Here, we present the evaluation of the three puzzle selection algorithms in the ESP game. In the simulation, we set $T = 10,000$ and $X = 6$; M denotes the total number of puzzles in the system. The results in Figure 5 show that, when M is small (say, smaller than a threshold M'), the three algorithms are comparable in terms of the system gain achieved. However, when M is larger than M', the system gain of OPSA remains consistent regardless of the changes in the values of M. In contrast, the system gain of FPSA and RPSA degrades as the value of M increases, and RPSA slightly outperforms FPSA when M is very large. More precisely, the threshold M' represents the minimal number of puzzles required to achieve the maximum system gain (i.e., $M' = N = T/r$). Since $T = 10000$ and $X = 6$, we know that $E[S] = 10.3353$ and $r = 31$. Therefore, $M' = 10000/31 \approx 321$ in this case. The results indicate that, when using the OPSA scheme, the ESP game must maintain at least a certain number of puzzles to achieve the maximum system gain[1]; otherwise, it will favor the RPSA and FPSA schemes because their performance is comparable to that of OPSA and they are easy to implement.

[1]Fortunately this is not a problem in general, since the number of the puzzles can be easily increased by adding new puzzles from the Internet.

6. Conclusion

In this paper, we have studied the ESP game, an emerging human computation system, and proposed a metric, called system gain, to evaluate the game's performance. Moreover, we argue that human computation games need to be *played with a strategy* in order to collect human intelligence in a more efficient manner. Based on our analysis, we propose and implement an Optimal Puzzle Selection Algorithm (OPSA) to provide guidelines for improving the ESP game. Using a comprehensive set of simulations, we have investigated the properties of the ESP game, and demonstrated that the proposed OPSA scheme substantially outperforms other schemes in all test cases. Moreover, the proposed analysis is simple and applicable to other ESP-like games, and the proposed puzzle selection strategy shows promise for use in the design and implementation of future human computation systems.

References

[1] Image Labeler. http://images.google.com/imagelabeler/.

[2] J. Howe. The rise of crowdsourcing. *WIRED Magazine*, 14(6), June 2006.

[3] A. Koblin. The sheep market: Two cents worth. Master's thesis, UCLA, 2006.

[4] J. Surowiecki. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday, May 2004.

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

[6] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *ACM SIGCHI*, 2004.