Human Computation Games for Commonsense Data Verification

Tao-Hsuan Chang, Cheng-wei Chan, Jane Yung-jen Hsu
Department of Computer Science and Information Engineering
National Taiwan University
doudi.tw@gmail.com, gattathree@gmail.com, yjhsu@csie.ntu.edu.tw

Abstract

Games With A Purpose (or GWAP) provide an effective way to collect data from web users. As the data grow over time, it becomes increasingly important to verify the data collected. This research explores the design of two games, *Top10* and *Pirate & Ghost*, for verification of sentences in our Common Sense Knowledge Base (CSKB) collected via the Virtual Pets game. *Top10* is a single-player game, in which the player attempts to guess the top answers to a given question. The frequency of common answers is used as the verification. *Pirate & Ghost* is a multi-player role-playing game, in which players cooperate to navigate to a target. The common routes are used to verify relations among concepts in the CSKB. This paper presents the experiments to evaluate the performance of each game, and shows how the games can be coupled to achiever higher efficiency and precision.

Introduction

Games With a Purpose(Von Ahn 2006) give rise to a solution to large-scale data collection, e.g. building a commonsense knowledge base. Players not only have fun but also contribute new data to the collection in the process. Unfortunately, malicious players, imprecise relations, and typing errors contribute to noisy data. We designed two human computation games, *Top10* and *Pirate & Ghost*, to verify the data in common sense knowledge base (CSKB). Top10 is Family-Feud-like game. Pirate & Ghost is a multi-player role playing game in the network of concepts from CSKB. User inputs to both games are used for verification.

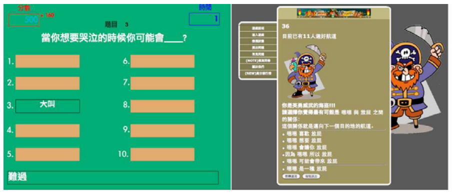
This paper introduces the game design of both *Top10* and *Pirate & Ghost*, and presents two experiments to evaluate their performance. The results showed that GWAP can be an effective tool for data verification.

Related Work

There are a couple of successful GWAP for collecting commonsense knowledge. The Virtual Pet Game (Kuo et al. 2009) is designed for players to interact with their pets by asking and answering commonsense questions based on fixed templates. Players help their pets with homework consisting of five answered questions. Previous study showed the game to have a high rate of new data collection, but a low rate of ranking collection and inadequate precision. Verbosity (von Ahn, Kedia, and Blum 2006) is a two-player game. The Narrator enters 5 words to describe a secret word, and the Guesser tries to identify the secret word with the least number of trials. The data collected by Verbosity also contain errors, such as IsA(bed, sleep) or HasProperty(read, book).

Game Design

Given the problems, it is essential to verify the correctness of data collected by GWAP. Each sentence in CSKB consists of two concepts linked by one relation. In this research, verification of knowledge tuples is broken into two parts. We designed the *Top10* game to verify the concepts and the *Pirate & Ghost* game to verify the relations in CSKB.



當你想要哭泣的時候你可能會____?

3. 大叫

難過

目前已有11人選好航道

你是英勇威武的海盜!!!
請選擇你覺得最有可能是 嗯嗯 與 放屁 之間的關係:
這個關係就是通向下一個目的地的航道。

- 嗯嗯 喜歡 放屁
- 嗯嗯 想要 放屁
- 嗯嗯 會讓你 放屁
- 因為 嗯嗯 所以 放屁
- 嗯嗯 可能會帶來 放屁
- 嗯嗯 是一種 放屁

Figure 1: Screenshot of Top10(L) and PG(R)

Top10

Top10 is a game for players to test the commonsense knowledge similarity between theirs and others. Game will give players one question with ten covered answers in each round. Players must input the top ten answers during 1 minute in our setting. When players match one of the top ten answers, their score will increase, and the answer will be revealed. When time is up, all top ten answers will be revealed. Players can rank the answers as good or bad top ten answer. After ten rounds, this game is over. Player's score is recorded on the ranking board. Top10 continuously collects answers and ranks from players' contribution and update the data in database, such that the top ten answers of each question in database will become more and more acceptable. Questions and their top ten answers are from assertions of CSKB and Top10 database. Specifically, an assertion includes three parts(concept → relation → concept). A question is formed if there are more than ten assertions that share same subject and relation but different objects. For example, a assertion ”Students don’t like homework” can be split to a question ”Students don’t like ____ ?” and an answer ”homework”. Data collection of Top10 can be divided into the data verification and collection of new commonsense. We call a input answer ”verified” if the answer is already in VP’s database. Otherwise it is ”unverified”. We also collect good or bad rank in the end of each round. The rank can verify the original rank in VP if the answer is ranked good in both game.

Pirate & Ghost

Pirate & Ghost (P&G) is a commonsense guessing role playing game(RPG) in commonsense knowledge network. A player plays as either a pirate or a ghost. The game flow is a synchronized round-based game. Whenever a new player join the game, he plays as a pirate with 3 lifes. If a pirate loses all life, he becomes a ghost. If a ghost rest 3 times, he will revive and become pirate again. In each round (1 minute), every player read a pair of concepts (”the location”), and decide what the correct relation (the ”passage”) between them is. There are total 37 passages. In the same time, the game page will notify players their role and give players hint says that pirates are going to select the correct passage and that ghost are going to select the incorrect passage. At the end of the round, game will sum the number of players select on each passage. If ghosts are more than pirates on a passage, all pirates lose 1 life, and all ghosts rest 1 time. Otherwise, nothing happens. Moreover, if number of pirates on a passage is small than 1/3 over total pirates, they will also lose a life (as pirates are fighting each others). Then pirates who lose no life will get a point, which can be shown on a score board. Finally, the game calculate each player’s new status, and announce the new round. The game start a new round by use the previous round location and another related concept which is immediately linked in common sense knowledge base(CSKB). The original relationships in CSKB are omitted in game. We collect pirates’ selections as correct relation predictions and ghosts’ selections as incorrect relation predictions. We also assume the passage most pirates chose is the true relation. Then we record the pair of concept and all the aggregated selection by the tuple (passage, role, number) into the game database.

Evaluation

The participants were recruited on PTT using virtual currency. In the Top10 study, 104 participants played a total of 1866 questions. There were 13624 verified answers, 14729 unverified answers and 9040 ranks collected. On the average, there were 15.03 verified answers and 4.84 ranks generated per question within one minute. In the P&G study, 14 participants played a total of 58 rounds, and 522 passages were collected. On the average, 9.00 passages were generated during each round of game play. After filtering out the training rounds, 442 passages remained. Among them, only 12.90%(57/442) overlapped with the assertions in CSKB. Given our focus on the verification of assertions in CSKB, only the overlapping data were considered for evaluation. We then invited 17 volunteers as the “commonsense experts” to vote whether each answer is true or false. Each assertion is voted till there are at least five votes for either true or false, which are then used as the ground truths in evaluating the game data.

Top10

The quality of player answers has a precision of over 90% among answers receiving good ranks from the volunteer “commonsense experts”. For answers with high frequency, the precision further increases to 93%. However, some answers may be ambiguous and cannot be easily ranked as good or bad. For example, the assertion “Teachers hate Students” is ranked both good and bad by different users. For ranks’ quality measurement, we examined the correlation between rank in Top10 and commonsense experinced users (experts). For good rank in Top10, 93%(609/656) are the same as experts. However , only 32%(104/152) bad rank in Top10 are the same as experts. It is likely that Top10’s players give bad ranks to acceptable but less relevant answers, while experts rank them good.

Pirate & Ghost

We use the votes as ground truth to compare the Pirate & Ghost(P&G) game data and CSKB data by calculating the precision. We take the assertion with higher “true” vote as true assertion and vice versa. Since there are no even votes, each assertion is either true or false. Then we use P&G passages and CSKB ranks as prediction, ignoring the false assertions. The precision of P&G pirate passages is 96.08%(49/51) whereas the one of CSKB good ranks is 86.96%(20/23). Moreover, we combine both CSKB and P&G record with correct prediction. Assertions predicted correct in both systems are all true(11/11). The intersection is above 1/5 size of P&G and about 1/2 size of Virtual Pets. Therefore, we can combine data of verification games with CSKB data to get the verified data.

We have presented two human computation games to verify the commonsense knowledge collected from another game. This work suggests that GWAP are effective as a tool for data verification.

References

Kuo, Y. L.; Lee, J. C.; Chiang, K. Y.; Wang, R.; Shen, E.; Chan, C. W.; and Hsu, J. Y.-j. 2009. Community-based game design: experiments on social games for commonsense data collection. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*.

von Ahn, L.; Kedia, M.; and Blum, M. 2006. Verbosity: A game for collecting common-sense knowledge. In *ACM Conference on Human Factors in Computing Systems (CHI Notes)*, 75–78.

Von Ahn, L. 2006. Games with a purpose. *Computer* 39(6):92–94.