Human Computation
AAAI Technical Report WS-12-08

Collecting Representative Pictures for Words: A Human Computation Approach based on Draw Something Game

Jun Wang and **Bei Yu**
School of Information Studies, Syracuse University
Hinds Hall, Syracuse, New York 13244

Abstract

This poster proposes a human computation approach to collecting representative pictures for words so that the collected pictures can efficiently and effectively convey the meaning of the words or concepts. A large collection of representative pictures can be used in text-to-picture communication systems, and may also be used to teach computers to learn what representative pictures are. We have developed a web application to help players of Draw Something, a popular social mobile game, search pictures for drawing inspiration while at the same time they implicitly help us collect representative pictures for words. Our preliminary result shows that the proposed approach has the potential to harvest Draw Something players for collecting desired data.

Introduction

"A picture is worth a thousand words." Visual representations have long been shown to increase human's attention, comprehension, and recall of linguistic texts. There are numerous studies in AI on translating pictures to texts (i.e. image understanding, recognition, or annotation), but much fewer studies on automatically translating texts to pictures (UzZaman, Bigham, and Allen 2011; Zhu et al. 2007; Mihalcea and Leong 2008). A key component of text-to-picture systems is to identify representative pictures that can efficiently and effectively convey the meaning of the concepts in the text. In this poster, a picture is called *representative* if it only carries necessary information for conveying the meaning (efficiency), and it includes sufficient information for conveying the meaning (effectiveness).

Existing approaches of identifying pictures for words usually use pictures from manually-created clipart libraries, images extracted from Wikipedia, or images retrieved from search engines. There are also studies on automatically generating representative images for words (Li et al. 2008; Zhu et al. 2007), or building a WordNet-like image ontology database (Deng et al. 2009).

Different from existing work, the main question we ask in this poster is: Can we design a system to harvest the *crowd* to collect representative pictures for words?

To answer the question, we have developed a web application to help players of Draw Something search pictures for use as drawing models while at the same time they implicitly help us collect representative pictures for words. Draw Something is a pictionary-like social game. In the first 6 weeks of its launching, the game was downloaded 30 millions times and generated over 3000 drawings per second. In the game, players are asked to draw a picture to represent a given word, and their partners will then need to guess the word from the drawing.

Our idea is that when a user of our tool spends a significant amount of time on a specific picture during the image search and browse process, we assume that the user might adopt the picture as a model for his drawing. Note that an important feature of Draw Something is that the game play is asynchronous rather than real-time, thus allowing a drawer to make a drawing for as long as he likes. We further assume that a picture, when chosen as a *drawing model* for a given word, is a representative one because (1) it is simple, easy to draw, minimizing unnecessary details, and so it can efficiently convey the meaning of the word, and (2) it must convey sufficient information for the guesser to figure it out.

To capture which picture is chosen by a user as a drawing model, we add a layer of opacity to all the pictures returned by our picture search function (see Figure 1). The opacity is set to such a level that those pictures are still viewable, but if one wants to see them sharply he may need to remove the opacity layer. We tell users that they can click a specific picture to focus on it (i.e. removing the opacity layer). By recording the time when a user clicks a picture and the time when he leaves the picture page, we can estimate how much time the user might spend on the picture.

Currently our image search function is based on the Bing search API, and for the sake of this preliminary study we only show users a fixed set of 9 pictures. The queries sent to Bing are formulated by adding a target word with a modifier "clipart."

Preliminary Results

Because of the time limitation of data collection, we only have some preliminary results here. The specific questions we aim to answer are: (1) Can the proposed approach be used to collect desired data? (2) Do users have common preferences on which pictures to choose?

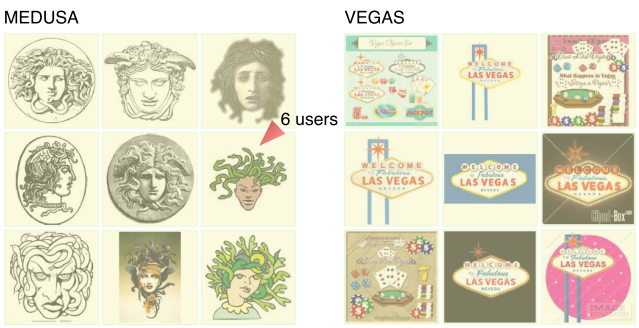

Figure 1: Pictures returned for a word are covered by a layer of opacity. When a user clicks a picture to reveal it, and then stays on the picture for at least 30 seconds, we assume that the user chooses the picture as his drawing model. In other words, the user implicitly votes for the picture. *Left panel:* All the 6 users independently chose the same picture for making a drawing for word "medusa." *Right panel:* 12 users have generated drawing models for word "vegas," distributed over 6 pictures with frequency distribution (3, 3, 2, 2, 1, 1).

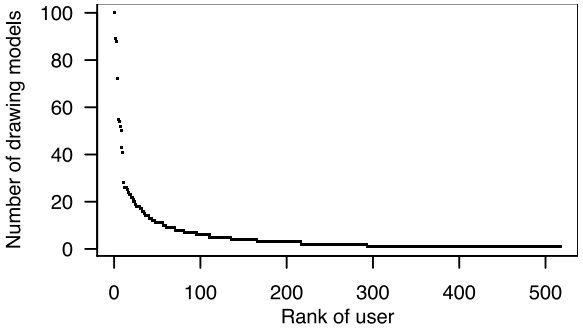

Figure 2: Power law distribution: Users are rank-ordered by the number of drawing models they generated. Among the 517 users, 10 of them each generated 40 or more drawing models.

Within two weeks of deploying the above approach in our web application, we have collected 2603 drawing models from 517 users (i.e. IP addresses) on 1084 words. When a user has been recorded to spend at least 30 seconds on a picture, we say that the user has generated a *drawing model*; in other words, the specific picture is referred to as a drawing model. Figure 2 shows how many drawing models were generated by each of these users. Note that during the image browse or focus process, a user may just "silently" choose a picture as a model without clicking any picture to remove its opacity layer. If this happens, we will fail to collect desired data from users, though we know that they do spend at least 30 seconds on the picture page. We found that there are 1989 such cases. Obviously, there is a trade-off between the effectiveness of data collection and the usability of the tool: if the opacity level is set to be too dark, we could collect more data from users but we may bring users bad experience. In our current setting, adding a layer of appropriate opacity to the retrieved pictures can engage over half of users (accurately speaking, 57% users) in providing data to us implicitly. In summary, this result shows that the proposed approach can satisfactorily be used to collect desired data from players of Draw Something.

Our approach to the second question is to find out the difference between the average "approval" rate of the favorite picture for each word and their expected "approval" rate if assuming no consensus among the players regarding picture choices. For each word, we calculate the average "approval" rate as the number of votes on the most popular choice divided by the total number of votes. For example, "medusa" attracted 6 votes, all for one picture (see Figure 1). Its approval rate is thus 100%. "vegas" attracted 12 votes, and two pictures tied with 3 votes for each. The approval rate in this case is thus 3/12=25%. However, the approval rate measure is not reasonable when the number of votes is very small. For example, "balloon" attracted only one vote, and the approval rate would be 100%. We choose to set the minimum number of votes to 6 in order to have enough number of cases for the one-sample t-test, which requires at least 30 cases to lift the assumption of normal distribution. In the end, we obtained 48 cases. Our null hypothesis H_0 is that users do not have consensus on which picture(s) to choose. If H_0 is true, the expected approval rate should be 1/9=0.11. We use SPSS to run the t-test. The average approval rate is 60.9%, standard deviation 19.7%. We set the significance level to 0.01. The one-sample t-test result shows the average approval rate is significantly higher than the expected value ($t = 17.546$, $p < .001$), therefore we rejected H_0 and support the claim that users do have consensus on choosing representative pictures. We then tested a stronger null hypothesis that the average approval rate is not higher than 50%. Again, the t-test result rejected the null hypothesis ($t = 3.825$, $p < .001$) and support the claim that on average the favorite pictures win more than 50% majority vote.

In conclusion, our preliminary results support the feasibility of using the proposed approach to collect representative pictures for words, and demonstrate that users tend to share common preferences on which pictures to choose.

References

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 248–255.

Li, H.; Tang, J.; Li, G.; and Chua, T. 2008. Word2Image: Towards Visual Interpreting of Words. In *ACM MM*, 813–816.

Mihalcea, R., and Leong, C. W. 2008. Toward communicating simple sentences using pictorial representations. *Machine Translation* 22(3):153–173.

UzZaman, N.; Bigham, J.; and Allen, J. 2011. Multimodal summarization of complex sentences. In *IUI*, 43–52.

Zhu, X.; Goldberg, A.; Eldawy, M.; Dyer, C.; and Strock, B. 2007. A text-to-picture synthesis system for augmenting communication. In *AAAI*, 1590–1595.