A Technique for Ranking and Visualization of Crowd-Powered Subjective Evaluations

Erika Gomi[1)], Yuri Saito[1)], Takayuki Itoh[1)], Mariko Hagita[1)], Masahiro Takatsuka[2)]
1) Ochanomizu University 2) The University of Sydney
{erika53, yuri}@itolab.is.ocha.ac.jp, {itot, hagita}@is.ocha.ac.jp, masa.takatsuka@sydney.edu.au

Abstract

We previously presented a crowd-powered digital contents evaluation system. This system shows a lot of pictures to the answerers and ask them to input the evaluations. It preferentially selects pictures which are predicted to be highly or poorly evaluated to the answerers, based on our assumption that high or poor evaluations are more informative results comparing with moderate evaluations. We have applied an interactive genetic algorithm in our system to select such pictures. This paper presents a technique for ranking and visualization for the evaluation results collected by our system. The presented technique calculates scores of all contents and uses for the ranking. Here, it may happen that some pictures are shown to no answerers while using our evaluation system. Our technique presented in this paper estimates the evaluation of such pictures shown to no answerers, and finally complete the ranking of all the pictures. The paper also presents the visualization tool for the ranking of pictures, and our experiment to demonstrate the effectiveness of our technique.

Keywords--- Evaluation, Ranking, Visualization

1. Introduction

Subjective evaluation results of commercial products and services are recently well published on the Internet. For example, Web sites for recommendation of restaurants or hotels publish average scores of customers. Or, Web sites for survey commercial products such as clothes or vehicles publish average scores of questionnaire results. Usually each of customers does not input the rates or all services or products for these Web sites. Most of them just input the rates for small number of services or products. In other words, crowd-powered inputs applied to such Web sites are useful for our daily life. We are addressing the development of a system to effectively collect such evaluations with small tasks of answerers.

We have presented an interactive evaluation system [1] for impression evaluation of digital contents. This system iteratively shows pictures to answerers and asks them to input evaluations for the pictures. The system applies an interactive genetic algorithm (iGA) so that we can select pictures predicted as highly or poorly evaluated. Generally it is more informative if we can suggest which contents are highly or poorly evaluated as the result of impression evaluations. Based on this concept, we aim to collect answers of contents predicted as highly or poorly evaluated. We expected this mechanism reduces tasks of answerers to obtain informative answers, and actually we demonstrated the effectiveness of the mechanism [1].

We found a problem of this system that some contents may be shown to no answerers, especially when large number of contents are prepared, or when small number of answerers participated. Therefore, it was difficult to make a complete ranking or all contents by our previous system. To solve the problem, we propose an algorithm to score all the contents based on the evaluation results, and a technique to estimate scores of contents which are shown to no answerers. We can make a ranking of all the contents by using these techniques after completing the crowd-powered evaluation tasks.

We created 1536 pictures of female faces for impression evaluation of appearance of women for the experiment of this study. In detail, we firstly generated 16 average female faces by blending real photographs of female faces. Then, we retouched their hair styles and make-ups by using commercial face retouch tools. We applied these pictures to our evaluation system and asked 30 participants to evaluate them. We then applied the techniques presented in this paper to score all the pictures and visualize the ranking. This paper introduces this experiment and discusses the effectiveness of the presented technique.

2. Related work

Crowdsourcing has been recently well-applied to construct collective intelligence. Especially it is useful for various academic and industrial fields which require subjective evaluations. Crowd-powered digital contents design techniques have been applied for Web design [2], 2D image generation [3] and 3D geometric modeling [4]. Koyama et al. [5] presented a more generic technique for optimization of digital contents design applying a crowdsourcing task. Our technique is different from

these existing techniques since our technique does not optimize parameters but assists the selection of contents.

Interactive genetic algorithm (iGA) has been applied to wide range of applications, such as image retrieval [6], music recommendation [7], and recommendation of cloth coordination [8]. Differently from these studies, our technique applied an extended iGA which explores the best and the worst solutions at the same time.

3. Evaluation System Applying an Interactive Genetic Algorithm

We have presented an interactive evaluation system which shows small number of contents to users and requests to input their evaluations [1]. We can construct the crowd-powered knowledge by collecting the evaluations of many users by using this technique.

During the development of this system, we supposed it should be efficient if we can collect high or poor evaluation of contents from small number of users, or small number of answers for each user. Based on this assumption, we developed an interactive contents evaluation technique which preferentially shows the contents which the technique predicts the user will highly or poorly evaluate, applying interactive genetic algorithm (iGA). Here, ordinary iGA just explores the highly adapted solutions, while our technique requires an algorithm which simultaneously explores highly and poorly adapted solutions. Therefore, our implementation of iGA applies an island model [9] to divide the individuals into two islands and separately explore highly or poorly scored contents respectively.

Following is the processing flow of the presented crowd-powered contents evaluation technique.

Step 1: Initialize Population

Select constant number (12 in our implementation) of contents as initial individuals randomly.

Step 2: Display

Show the contents to users.

Step 3: Evaluation

Request the users to input the subjective evaluation for the contents. Our implementation provides three button widgets corresponding to "Good", "Soso", and "Bad", and requests the users to press one of them.

Step 4: Selection and Immigration

Collect individuals evaluated as "Good" to the island of "Good". Similarly, collect individuals evaluated as "Bad" to the island of "Bad". Dismiss other individuals evaluated as "Soso".

Step 5: Crossover

Generate new generation of the individuals in the two islands respectively.

Step 6: Mutation

Randomly apply the mutation for the diversity of individuals.

Step 7: Termination

Stop the iteration if it satisfies pre-defined conditions. Our current implementation just terminates if the sequential number of the current generation exceeds the pre-defined number.

4. Ranking of Evaluation Results

This section presents a technique to score all the contents from the evaluation results constructed from the answers collected by the interactive evaluation system introduced in the previous section. This section also presents a technique for estimation of the evaluation for contents which are shown to no answerers.

4.1 Estimation of evaluation for contents which are shown to no answerers

It may happen that some contents are shown to no answerers while using the interactive evaluation system introduced in the previous section. On the other hand, we need to score all the contents from the evaluation results, in order to complete the ranking of all contents.

To solve this problem, we developed a technique to estimate the evaluation of such contents applying Self-Organizing Map (SOM). SOM is an unsupervised neural network used for mapping multi-dimensional vector items onto low-dimensional spaces. Data items which have similar vectors are closely placed in the low-dimensional space.

Following is the processing flow of our technique to estimate the evaluation applying SOM, where

- N_c is the number of attributes of the contents,
- $\{a_1, \ldots, a_{NC}\}$ is the attribute values of a content,
- g is the number of answer of "Good",
- s is the number of answer of "Soso", and
- b is the number of answer of "Bad".

1. Divide the contents into training and test groups. Training group consists of contents evaluated by one or more answerers. Test group consists of contents evaluated by no answerers.
2. Form vectors corresponding to contents in the training group. We form a (N_c+3)-dimensional vector of a content as $\{a_1, \ldots, a_{NC}, g, s, b\}$.
3. Generate a SOM from all vectors corresponding to all contents in the training group.
4. Estimate g, s, and b of the contents in the test group. We form a N_c-dimensional vector of a content as $\{a_1, \ldots, a_{NC}\}$, and extract similar vectors in the training group mapped onto the SOM. This process then calculated weighted average values of g, s, and b from the extracted vectors. These values are used as the estimated values of the current content in the test group.

4.2 Scoring and ranking of the contents

After estimating evaluations of all contents in the test group, the technique calculates scores of all contents. We calculate the scores based on the ratio of numbers of "Good", "Soso", and "Bad".

We calculate the score of the i-th contents $Score_i$ by the following equation:

$$Score_i = \sum_{j=0, j\neq i} \frac{g_i(s_j + b_j) + s_i b_j}{sum_{ij}} - \frac{g_j(s_i + b_i) + s_j b_i}{sum_{ij}}$$

where
g_i is the number of answer of "Good" for the i-th content,
s_i is the number of answer of "Soso" for the i-th content,
b_i is the number of answer of "Bad" for the i-th content,
and $sum_{ij} = (g_i + s_i + b_i)(g_j + s_j + b_j)$.
After calculating the score of all contents, we simply make a ranking of the contents in the descending order of the scores.

5. Example

5.1 Dataset of female face images

We used the set of images of women's faces introduced in [1] as an example dataset. We created these face images by the following process. We firstly took face pictures of 18 twenties women, and generated intermediate images by applying a morphing technique. As a result, we generated 16 types of intermediate face images as the combination of the following features.

- Length of the face: "long" or "short".
- Form around the chin: "thin" or "round".
- Impression of eyes: "bright" or "thin".
- Impression of nose: "thin" or "round".

Then, we applied a makeup simulation service (SHISEIDO Beauty check point makeup) and a hair style simulation service (Hairstyle Simulator "ChouChou") to generate more variety of face images. We applied the combination of the following features for the face image synthesis.

- Makeup type: "fresh", "cute", "cool", or "elegant".
- Length of hair: "long", "medium", or "short".
- Bangs: "with" or "without".
- Form of hair: "straight" or ``waved".
- Color of hair: "brown" or "black".

We generated 1536 face images as a result. Figure 1 shows examples of face images. These face images are coded as 9-dimensional vectors and applied to iGA implemented in our interactive evaluation system.

5.2 User experiment

We had a user experience with 30 female participants using our implementation of the evaluation technique for appearance of women. All female participants are university students majoring computer science. The following is the setting of the iGA in our experience.

- Total number of face images: 1536
- Number of individuals in a generation: 12
- Termination condition: 20 generations
- Crossover ratio: 1.0
- Mutation ratio:
  - if n_{soso}<4 : 0.05
  - otherwise : 0.05(n_{soso}-2)
  - where n_{soso} is the number of images which a user evaluated as "Soso" in the previous generation.

Average number of face images showed to participants was 169, which means that each of the participants evaluated just approximately 10% of the contents. Detail of the result is described in [1].

Figure 1 Examples of synthesized face images.

In this experiment, 54 images were shown to no answerers. We estimated the evaluations of these images by applying SOM. Then, we calculated the scores of all images. Figure 2 shows the distribution of the scores. Here, scores are assigned to the vertical axis, and all the images are sorted according to their scores and aligned along the horizontal axis. Figure 2(Upper) shows that scores of unevaluated 54 images were zero. Figure 2(Lower) shows that scores of all images look smoothly varied.

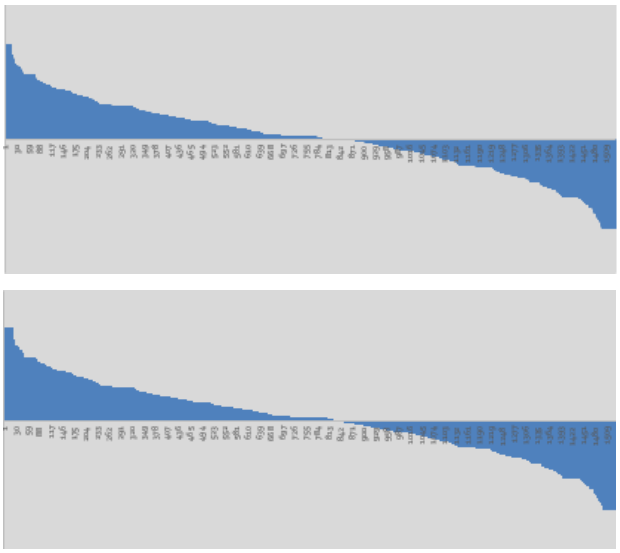

Figure 2 Distribution of scores. All images are sorted according to their scores and aligned along the horizontal axis. (Upper) Before estimating evaluations of images shown to no answerers. (Lower) After estimating evaluation.

6. Visualization Tool

We developed a visualization tool for the ranking result. Figure 3 shows a snapshot of the visualization tool.

Figure 3(Upper) shows an initial display of this visualization tool. The left side of the window features buttons used for interactive specification of attributes of contents. While visualizing the dataset introduced in Section 5, the buttons are used to select the shapes of faces and face parts, makeup types, and hair styles. The right side of the window displays pictures of the contents in the order of the scores. A color bar featuring blue, green and red parts depicts the ratio of evaluations of "Good", "Soso", and "Bad".

If a user selects some attributes by operating the buttons, pictures corresponding to the selected attributes are displayed in the right side of the window. Otherwise, all pictures are displayed. Also, we can zoom up to a particular content by a clicking operation, as shown in Figure 3(Lower). Users can interactively explore what types of contents are highly or poorly evaluated by using this visualization tool. We can focus on highly evaluated contents by displaying pictures in the descending order. Or, poorly evaluated contents are firstly displayed by selecting the ascending order.

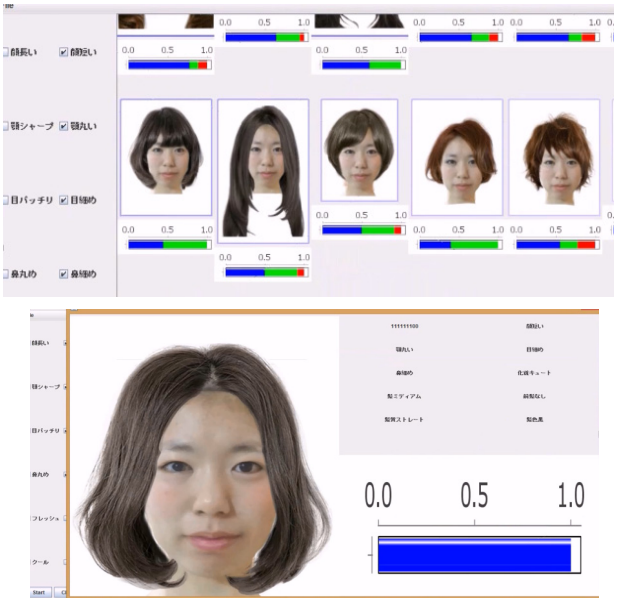

Figure 3 Snapshot of the visualization tool. (Upper) Display of contents in the order of scores. (Lowe) Focus on a particular content.

7. User Feedback

We tested the ranking results with 12 female participants. The participants were university students majoring computer science, who did not participate the experiment introduced in Section 5.2. We used the dataset of female face images introduced in Section 5.1 for this test.

Naturalness of the score estimation by SOM.

We created two small datasets including 24 face images. In each dataset, three of the images were shown to no answerers by the interactive evaluation system and therefore their scores are estimated by SOM. In other words, the images could be divided into 21 images in the training group and 3 images in the test group. We showed each of the datasets to the participants and asked them to guess which images were applied to SOM as ones in the test group. For both datasets, just one participant correctly selected one of the images whose score was estimated by SOM. Other eleven participants could not guess any of the images in the test group. This result suggests that the score estimation results were so natural and therefore participants could not correctly guess.

Inappropriateness of ranking result

We asked participants to find partial sequences of face images in the ranking result which they disagree the ranking. As a result, eight participants mentioned one or more sequences of face images. However, no same sets of images are mentioned by multiple participants. This result suggests that the ranking result may contain partial disagreement based on users' preferences; but it did not contain any portions which many participants disagree.

Operation of attributes

We asked participants to play with the visualization tool while freely selecting attributes including shapes of faces and face parts, makeup types, and hair styles. We then asked them to answer how they selected the attributes. Following is the statistics of the selection of participants:

- 8 participants selected shapes of faces and face parts closer to ones of participants' own faces and face parts.
- 5 participants selected attributes just based on their intuition and interests.
- 3 participants selected face parts which the participants were worrying about.

The result suggests the visualization tool would be used for the decision making of makeups and hair styles, because many of the participants selected the attributes based on their own appearances.

8. Conclusions

This paper presented a visualization tool for sets of digital contents which are completely scored and ordered based on evaluation results. This study is based on our previous development on an interactive evaluation system applying iGA. We proposed a technique for scoring contents based on the evaluations by using our interactive evaluation system, and a technique for estimating the evaluations of contents which are shown to no answerers. This paper then presented a visualization tool to interactively explore highly or poorly scored contents. We also introduced an example experiment and user feedback to demonstrate the effectiveness of the presented technique.

As future work, we would like to improve the implementation of iGA to realize more reliable and quick

user evaluation. Also, we would like to apply more variety of digital contents as well as female face images.

Acknowledgement

This work has been partially supported by Japan Society of the Promotion of Science under Grant-in-Aid for Scientific Research.

References

[1] E. Gomi, Y. Saito, T. Itoh, Visualization of Crowd-Powered Impression Evaluation Results, 18th International Conference on Information Visualisation (IV2015), 89-94, 2015.

[2] K. Reinecke, K. Z. Gajos, Quantifying Visual Preferences around the World, ACM SIGCHI Conference on Human Factors in Computing Systems, 2014.

[3] Y. Gingold, A. Shamir, D. Cohen-Or, Micro Perceptual Human Computation for Visual Tasks. ACM Transactions on Graphics, 31(5), 119:1-119:12, 2012.

[4] A. Secord, J. Lu, A. Finkelstein, M. Singh, A. Nealen, Perceptual Models of Viewpoint Preference, ACM Transactions on Graphics, 30(5), 109:1-109:12, 2011.

[5] Y. Koyama, D. Sakamoto, T. Igarashi, Crowd-powered Parameter Analysis for Visual Design Exploration, ACM Symposium on User Interface Software and Technology, 65-74, 2014.

[6] H. Takagi, S. Cho, T. Noda, Evaluation of an IGA-based Image Retrieval System Using Wavelet Coefficients, IEEE International Conference on Fuzzy Systems, 1775-1780, 1999.

[7] Y. Saito, T. Itoh, MusiCube: A Visual Music Recommendation System featuring Interactive Evolutionary Computing, Visual Information Communication and Interaction Symposium, 2011.

[8] S. Cho, Towards Creative Evolutionary Systems with Interactive Genetic Algorithm, Applied Intelligence, 16(2), 129-138, 2002.

[9] D. Whitley, S. Rana, R. B. Heckendorn, The Island Model Genetic Algorithm: On Separability, Population Size and Convergence, Journal of Computing and Information Technology, 7(1), 33-47, 1999.