Mining Human Attractiveness from Chinese Dating Game Show
If You Are The One

Xuelei Hu, Xiao He, Junwei Zhang
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing 210094, P.R. China
Email: xihu@njust.edu.cn

Bo Yuan
Division of Informatics, Graduate School at Shenzhen
Tsinghua University
Shenzhen 518055, P.R. China
Email: yuanb@sz.tsinghua.edu.cn

Abstract—Matchmaking is an everlasting topic in human society. In this paper, we focus on human attractiveness in matchmaking for the new generation of individuals in China based on If You Are The One, a famous dating game show in China. We collect a comprehensive contestant dataset, including numeric, text and image attributes. We investigate the influence of various factors on human attractiveness using correlation analysis and average face technology. Experimental results show that attractiveness is closely related to job and age, and facial features are more important for female than for male. A visualization of the matching network is also given to further explore some interesting phenomena.

Keywords—social behavior; human attractiveness; data mining; social science;

I. INTRODUCTION

With the rapid development of information technology, computational social science [1] has attracted increasing attentions from various disciplines in recent years. Many effective techniques have been developed for representing, modeling, analyzing, understanding and managing human behaviors and social characteristics [2], [3], [4]. Matchmaking or human mate selection is an everlasting topic in human society. Researchers have been seeking to identify mate choice preferences for many years[5], [6]. Most existing approaches focus on characteristics such as age, race, religion, income, and education [6].

In this paper, we used a popular Chinese dating game show If You Are The One to explore the mate selection behavior of the younger generation in China and focused on mining relevant factors of human attractiveness. If You Are The One is inspired by the Australian game show Taken Out and is produced by Jiangsu Satellite Television (JSTV). It was first broadcast on January 15, 2010 and its episodes are also widely distributed online. Due to its unique approach to dating and light-hearted style, as well as its humorous and open conversations, it has been a rating success in China and is now the highest-rated show for JSTV. It also attracted attentions from academics and foreign media and some studies have been conducted from the perspective of communication and psychology [7], [8].
II. DATA CONSTRUCTION

A. Data Collection

We collected the data from 100 episodes of If You Are The One, which were shown on JSTV from 3 Dec., 2011 to 18 Nov., 2012. The original dataset consists of two parts: 1) the profile of male and female contestants, such as name, age, occupation, photo, released on the official website (http://www.ifcwr.com/); 2) specific information from the show that was collected from the online videos.

The profile of each female contestant consists of the following 12 factors.

1) **ID**: The unique number assigned to each contestant.
2) **Name**: The name of the contestant.
3) **Photo**: A 550×730 color photo of the contestant. An example is shown in Fig. 2.
4) **Age**: The age of the contestant.
5) **City**: Current residence.
6) **Job**: Current occupation.
7) **Children**: Attitude on having kids in the future family.
8) **Previous Relationship**: Fact about previous relationships, such as “having three relationships before” or “having been single for a year.”
9) **Perspective on Consumption**: The viewpoint on consumption.
10) **Living with Parents-in-Law**: Whether the contestant minds living with her future parents-in-law.
11) **Parents**: The situation of the contestant’s parents including their occupation and marriage situation.
12) **Dislike**: The shortcomings of a male that the female contestant cannot tolerate most (e.g., being selfish).

The profile of a male contestant includes 6 factors: ID, Name, Photo, Age, City and Job, defined in a similar way as female contestants.

For the second part, we asked a group of volunteers to watch the episodes online and recorded some important factors. To help understand those factors, we first briefly introduce the rules of the show. In each episode, twenty-four female contestants stand in a line, each atop a podium with a light on. They face a single man who initially chooses one of them as the “dream girl” in secret. The male contestant has the opportunity to use 2 to 3 video clips to reveal additional personal information such as occupation, interests, love history and friends’ opinions. After his first appearance and after each video clip, if a girl does not like the man, she will turn the light off. During this period, the contestants, guests and host may often have some improvised dialogues to warm up the atmosphere. If all 24 lights go off, the man loses. If a single light remains after the man’s introduction, the man may choose to accept the woman or give up. If two or more lights remain, the man has the option to invite two women on stage. In this case, he can also invite the girl who he secretly selected at the beginning (even if her light is off). With up to three women on stage, the man will ask some questions and then make his final choice. If the girl accepts, they walk towards each other and head off the stage to start a relationship in the real life.

We collected the following 4 factors of each male contestant, which we believe are closely related to attractiveness.

1) **Dream Girl**: The girl chosen as the dream girl (most attractive to him at the beginning).
2) **Number of Lights (first)**: The number of remaining lights at his first appearance (first impression of female contestants).
3) **Number of Lights (last)**: The number of remaining lights at the end.
4) **Result**: Whether the male contestant successfully starts a relationship with a female contestant.

B. Data Preprocessing

Many factors collected above were texts or images and we will explain how to extract numerical features from them.

1) **Job**: We divided all original jobs into six classes: 1). Physical Worker, such as carpenter, cook, driver, mechanic, repairmen, farmer, and builder; 2). Technicians, such as scientific researcher, professor, engineer, computer programmer, system analyst; 3). Artist, such as actor, sculptor, architect, singer, composer, poet, and dramatist; 4). Manager, such as project manager, sales and marketing manager, government official, businessman; 5). Educator, such as teacher and social worker; 6). Clerk, such as secretary, accountants, administrative assistant, librarian, teller and typist.
2) **Children**: There were three possible situations: 0 for no plan to have children; 1 for one child; 2 for at least two children.
3) **Previous Relationship**: We split this factor into three new features: 1). Number of relationships before; 2). Marital status, 1 for married before and 0 for never married before; 3). Time of being single, showing the number of years during which the contestant has been single. For instance, 2-0-1 means that the contestant had two relationships before, never married, and has been single for one year.
4) Perspective on Consumption: We grouped this factor into three categories: 1 for hard-working and thrifty, 2 for moderate (there is no need to be too thrifty but should not spend without restraint either) and 3 for extravagant.

5) Living with Parents-in-Law: We classified this factor into three categories: 1 for unwilling to live with parents-in-law, 2 for do not mind living with parents-in-law and 3 for neutral attitude.

6) Parents: This factor was split into four new features: 1). Whether her parents are alive; 2). Whether her parents are divorced; 3). Her fathers occupation; 4). Her mother’s occupation.

7) Dislike: We listed 10 shortcomings that were most often mentioned in the show: 1). Selfish; 2). Lazy; 3). Playing the field; 4). No ambition; 5). Dishonest; 6). Problematic character; 7). Bad temper; 8). No sense of responsibility; 9). No family filial piety, 10). Bad personal habits. Thus we used a 10 dimensional Boolean vector to describe this factor. For example, (1 0 0 1 0 0 0 0 0 1) means that the shortcomings that this contestant cannot endure are being selfish, having no ambition and having bad personal habits.

As to the image data, we extracted 58 landmark points of a face as the geometric features. Firstly, we used Active Appearance Models (AAM) to automatically detect facial landmarks [9]. In cases where the automatic detection failed, we asked some volunteers to mark the images manually. Secondly, we converted the images to gray scale and normalized the landmarks according to the coordinates of eyes. Fig. 3 illustrates the preprocessing procedure in which the shift and scale variances are removed.

C. Attractiveness Score

Attractiveness plays an important role in loving relationships. There are many kinds of ratings that can be possibly used most of which require human judgments. Instead, we directly used the information from the show itself to measure the attractiveness of a contestant. A major advantage is that it can reflect the true feelings of other people on the show. For each female contestant, we counted the total times that she was selected by male contestants as the dream girl, noted by $N_{\text{dream}}$, and the total number of episodes that she participated, noted by $N_{\text{total}}$. Her attractiveness score was then calculated as:

$$A_{\text{female}} = \frac{N_{\text{dream}}}{N_{\text{total}}},$$

(1)

For each male contestant, we used the information of Number of lights (first), noted by $N_{\text{first}}$, and Number of lights (last), noted by $N_{\text{last}}$ to obtain his attractiveness score:

$$A_{\text{male}} = \frac{N_{\text{first}} - \bar{N}_{\text{first}}}{\sigma_{\text{first}}} + \frac{N_{\text{last}} - \bar{N}_{\text{last}}}{\sigma_{\text{last}}},$$

(2)

where $\bar{N}_*$ and $\sigma_*$ indicate the mean and the standard deviation of $N_*$ respectively.

III. RELEVANCE ANALYSIS

With the contestant dataset, we explored the relationship between attractiveness and various attributes. We used IBM SPSS Statistics 19 to conduct the relevance analysis including data frequency analysis, correlation analysis, and regression analysis. There were two types of variables: interval scale variables (e.g., age), and nominal scale variables (e.g., job). We conducted correlation analysis for interval scale variables and One-Way ANOVA analysis for nominal scale variables.

A. Correlation Analysis

Correlation refers to the strength of relationship between two variables. A strong correlation means that two or more variables have a strong relationship with each other while a weak correlation means that the variables are hardly related. The most widely used correlation coefficient is Pearsons correlation, which is also referred to as linear or product-moment correlation. Based on a sample of paired data $(x_i, y_i)$, the Pearson correlation coefficient is given by:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

(3)

where $r_{xy}, \bar{x}, \bar{y}, x_i, y_i$ are the sample correlation coefficient, sample mean, and the ith sample, respectively. Table I shows the result of correlation analysis between attractiveness and two interval scale variables.

From Table I, it is clear that item Number of relationships before had no obvious relationship with attractiveness, which
indicates that male contestants did not mind too much about the number of previous relationships of female contestants. By contrast, there was a strong correlation between attractiveness and Age. In other words, age is a critical factor when a male chooses his potential girlfriend and, given the negative coefficient value -0.166, the general trend is: the older the age of a female, the less her attractiveness.

B. One-Way ANOVA analysis

As to the nominal scale variables, we conducted One-Way ANOVA to check whether the values of these variables can significantly change the attractiveness. In Table II, variables with smaller significance values have more impact on attractiveness. We can see that Job and Marital status had the most significant impact on attractiveness. In the meantime, City was the least important factor for attractiveness. These findings show that male contestants tend to prefer certain occupations and also care about whether the female contestants have been married before. On the other hand, location seems to be not that important and people are willing to relocate if necessary.

IV. FACIAL ATTRACTIVENESS ANALYSIS

"Love at First Sight" is common in all cultures around the world and beauty is well known to be a key factor of attractiveness. In this section, we mainly focused on the impact of facial features on attractiveness.

A. Average Face Attractiveness

A basic feature of human cognition is the creation of prototypes. In facial beauty analysis, the average face of beautiful faces is often used as the beauty prototype. Here we calculated the average faces for the most attractive female and male contestants using average face techniques [10].

Fig. 4 shows the average faces of 80 attractive female and male contestants. In general, the average faces often receive higher attractiveness ratings than individual faces [11].

B. Visual Attributes Analysis

Visual attributes play an important role in human cognition. For example, people often assess and describe the attractiveness by measuring various distances on faces [12]. In our study, we extracted the following popular features: eye size, mouth size and shape of chin, according to normalized landmarks. Eye Size: For the left eye, we calculated the area of the quadrilateral made up by points 22, 24, 26, 28 (see Fig. 5). The size of the right eye was calculated in a similar way and the bigger one was selected as the eye size. Mouth Size: The mouth size was defined as the distance between point 44 and point 48. Chin: The shape of chin was defined by the angle of the line connecting point 1 and point 6.

In our experiments, we used the Pearson product-moment correlation coefficient to measure the correlation. Table III shows the correlation analysis result for female contestants in terms of eye size, mouth size and the shape of chin.

By contrast, we found that the facial features of a male are less important for his attractiveness. Fig. 6 shows the relationship between attractiveness and eye size for male
Table III

<table>
<thead>
<tr>
<th>Features</th>
<th>Pearson Correlation</th>
<th>Significance (2-tailed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eye size</td>
<td>0.141</td>
<td>0.118</td>
</tr>
<tr>
<td>Chin</td>
<td>-0.071</td>
<td>0.433</td>
</tr>
<tr>
<td>Mouth size</td>
<td>0.236**</td>
<td>0.008</td>
</tr>
</tbody>
</table>

**indicates strong correlation at significance level 0.01.

Figure 6. Relationship between eye size and attractiveness for male contestants.

contestants, which was almost uncorrelated. Consequently, compared with female, it is much harder to judge the attractiveness of a male from facial features.

V. NETWORK VISUALIZATION

A. Topology of Matching Network

Here we assume that if a male contestant chooses a female contestant as his dream girl at the beginning of the game or starts a relationship with her at the end then he loves the female, whereas a female contestant loves a male contestant if she accepts him at last. Fig. 7 shows the matching network. Each node stands for one male or female contestant. If A loves B then there is a directed edge from node A to node B. If a female node has large in-degrees, it means she is very popular. According to the visualization graph, most female contestants have been selected as dream girl several times before they find their true love. Furthermore, over 50% female contestants have been selected as dream girl.

B. Analysis of Edges

There are many ways to assign a weight to each edge in the graph. Here we consider the age difference between male and female contestants, which plays an important role in Chinese traditional family values.

Histograms in Fig. 8 show the distributions of age differences, which both are nearly Gaussian. It seems that being 3 years older than the female is most common for the male. A significant downturn appears if the male is 7 years older or 1 years younger than the female.

VI. CONCLUSION

In this paper, we investigated the attractiveness in human mate selection using data mining and pattern recognition techniques. To the best of our knowledge, we constructed the first comprehensive contestant dataset based on If You Are The One. We illustrated how characteristics such as age, occupation and facial features affect the mate choice. Moreover, we compared the difference between male and female in terms of attractiveness. We believe that the presented work can be inspiring for other studies on attractiveness and may help better understand Chinese matchmaking culture. Further improvements can be achieved by improving the quality
Figure 8. Distributions of age differences between each male contestant and (a) his dream girl and (b) the female contestant with whom he starts a relationship.

and quantity of the contestant database and conducting deeper analysis on the relationships between attractiveness and the profile of contestants.

ACKNOWLEDGMENTS

This work was supported in part by Ministry of Human Resources of China via the Science and Technology Activities Grant for Returned Scholars, a project funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD), and the National Natural Science Foundation of China (NSFC) (No. 60905030).

REFERENCES


[8] P. Huang, “Analysis the success of if you are the one from the perspective of communication,” Ke Xue Tan Suo, pp. 120–121, January 2011, in Chinese.


