# Text and Image Data Mining Based on Acquaintance Social Network

Hong Yuan
yh@smail.nju.edu.cn

Kuo Chi
ck@smail.nju.edu.cn

Yuhong Luo
lyh@smail.nju.edu.cn

Fan Yu
yf@smail.nju.edu.cn

## Abstract

*Wechat moments and QQ zone are different from public online communities like Weibo, Zhihu and Baidu tieba for they are built on the basis of friends, relatives and classmates who are familiar with each other. Therefore, we call this type of community an acquaintance social network , which is opposed to a stranger social network. In the acquaintance social network, expect for online communication, there often exists more direct offline contact, which makes the communication methods of acquaintance social network and the stranger social network different. With the rapid development of the Internet, ordinary users need to establish a good impression in the social network of acquaintances, and send moments to attract the attention of friends. Based on the real data collected in the QQ zone, this paper uses the heat to measure the popularity of each "moment", and thus models and analyzes how to produce more popular "moments" in the acquaintance social network, and proposes a method for constructing user portrait based on text, image and other data published by user in acquaintance social network.*

**Keywords:***acquaintance social network, natural language processing, image processing, user portrait*

## 1. Introduction

Chinese online communities can be roughly divided into two categories. One is the stranger social networks represented by Weibo, Zhihu, and Baidu tieba, in which the contents published by each user can be viewed by most users casually, regardless of whether the users know each other. Therefore, the contents are usually extensively beautified and modified, and the authenticity is difficult to judge. The other one is the acquaintance social networks ,like Wechat moments and QQ zone, in which the contents published by each user can be viewed by online friends who are classmates, colleagues, relatives and friends of the user. Therefore, the contents in such communities are closer to the users real thoughts and personality. By constructing user portraits through these contents, we can distinguish user groups better for the purpose of precision marketing.

In the meanwhile, with the rapid development of the Internet, social networks are more closely connected with people's daily communication, and ordinary users also urgently need to create a good image in the acquaintance social network by posting some modified and beautified "moments". However, the way to obtain heat in the stranger social network does not work in the acquaintance social network because the users receiving the "moments" usually has real contact and understanding of the user, such as self-portraits that are completely inconsistent with the real person are more likely to cause dislike of acquaintances.

Therefore, this paper models the relationship between content and popularity generated by acquaintances' social networks through machine learning and other methods, and explores what kind of content is more popular in acquaintance social networks. This paper measures the degree of popularity by heat, which is a value based on pageviews, clicks, and comments, which are defined in Concepts and Definition.

This paper takes the moments in QQ zone as an example to quantify and analyze each moment from the following indicators: content of text in moments, mood of text in moments, quality and beauty of images in moments, content of images in moments and the time when the user post the moments. Also, these metrics will also be used to automatically build a user portrait for each user.

## 2. Concepts and Definition

### 2.1. Data Source

The data are collected from QQ zone from June 10, 2014 to the present and all collected users have clearly understood and agreed to the study. The number of moments which are recorded as D is 14,462. After data cleaning, every moment is showed in Table 1.

Table 1. moments data fields

| fields | description | type |
|---|---|---|
| tid | unique identifier for each dynamic | string |
| like_num | count of approval | int |
| prd_num | count of pageviews | int |
| cmt_num | count of comments | int |
| cmt_total_num | total number of comments | int |
| time_stamp | timestamp when the moment is posted | int |
| content | textual content of each moment | string |
| uin_list | list of users who like this moment | json |
| cmt_list | list of users who write comments | json |

### 2.2. Definition of Heat(E)

Let the average number of each moment in D be m, the amount of praise is n, and the pageview amount is h, then the conversion relationship between comment volume and praise amount is a, and the conversion relationship between the amount of praise and the pageview is b:

$$a = n/m, \tag{1}$$

$$b = h/n, \tag{2}$$

e is the entropy of each moment:

$$e = m_i * a * b + n_i * b + h_i, \tag{3}$$

in which, $m_i$, $n_i$ and $h_i$ are number of comments, amount of approval and amount of pageviews of the $i^{th}$ moment respectively.

Thus heat of each moment E is defined as the normalized value of entropy e:

$$E_i = (e_i - e_{min})/(e_{max} - e_{min}), \tag{4}$$

in which $e_{max}$ represents the maximum entropy of the user who produced the $i^{th}$ moment, and $e_{min}$ represents the minimum entropy.

Thus, under the assumption that the number of friends of each user is changeless, the size of E is independent of the number of friends.

## 3. Implement

### 3.1. Text Content Classification Based on RNN

Text types in the acquaintances social networks tend to be short, fragmented, and often have special symbols with ambiguous semantics, resulting in complex and diverse semantics. In this paper, we have labeled and constructed a data set containing more than 7,000 texts, and used the Recurrent Neural Network model for character level training and learning, and the final F1 reached 0.83.

Table 2. text types

| id | type | description | example |
|---|---|---|---|
| 1 | Tourism and Sports | Including outing, outdoor or indoor sports | 真正的川藏线，从现在才开始 |
| 2 | Love and Family | Anything related to love and family, including anything that mentions girlfriend, boyfriend and family members | 啊……以后就不要找我要我女票的照片了 |
| 3 | Learning and Working | Anything related to study or work, including anything that mentions school, class, company, work, classmates, colleagues | 国际周让我深深的体会到什么叫“只要选课选得好……” |
| 4 | advertisement | Including advertisements for likes, requests for forwarding, selling things, or various micro-businesses | 亲测，这里装修，背景音乐，饮料价格，整体氛围都超nice，二基楼福利 |
| 5 | Daily Life | Trivial things in life other than those in 1, 2, 3 and 4, such as dressing, eating, accommodation, etc | 不如吃茶去 |
| 6 | Others | The text is too short or there is no text or the text does not contain Chinese | http://url.cn/WqvxQE |
| 7 | Insights on Life | A remark about life or ideals, which is usually unintelligible or a text with strong emotion | 失落的岁月得到的美好垃圾的脖子 |

Table 3. labeled data structure

| field | description |
|---|---|
| label | text type |
| content | text content |

**Construction of Dataset** 7 text types are established from the randomly sampled 7,230 text data from the original data set D, which is shown in Table 2 According to the seven types above, we labeled each of the text and get the data composed like that in Table 3: Finally, the number of each type of data is shown in Table 4: Because the total

Table 4. number of each type

| type | count |
|---|---|
| 1 | 440 |
| 2 | 229 |
| 3 | 1038 |
| 4 | 185 |
| 5 | 3051 |
| 6 | 1510 |
| 7 | 777 |
| total | 7230 |

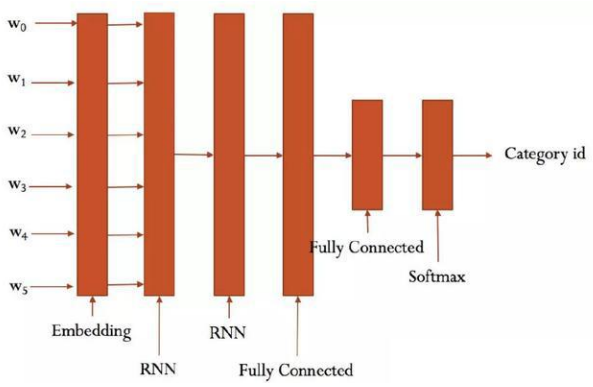


Figure 1. model framework

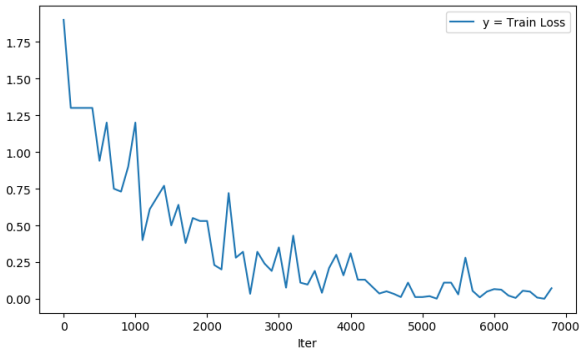


Figure 2. loss function

number is small, the training set, validation set, and test set are divided according to the ratio of 7:3:3.

**Training with RNN** The model framework is shown in Figure 1 The RNN is implemented by tensorflow and contains two hidden layers, and each layer contains 128 neurons. The convergence process of loss function and accuracy is shown in Figure 2 and Figure 2: The final test result is shown in Table 5 The confusion matrix is shown in Table 6

**Result Analysis** The final f1 is not very good and may

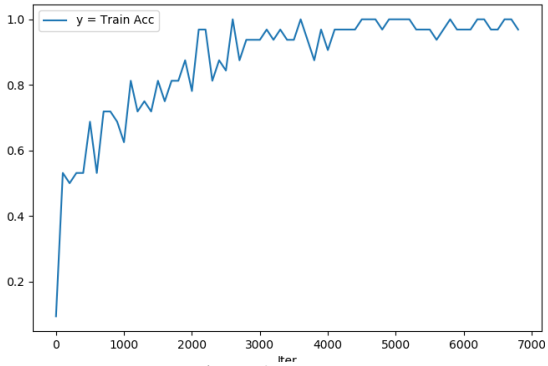


Figure 3. accuracy

Table 5. final test result

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| Tourism and Sports | 0.74 | 0.83 | 0.78 | 132 |
| Love and Family | 0.77 | 0.76 | 0.77 | 76 |
| Learning and Working | 0.8 | 0.82 | 0.81 | 308 |
| Advertisement | 0.74 | 0.67 | 0.7 | 55 |
| Daily Life | 0.86 | 0.87 | 0.86 | 919 |
| Others | 0.88 | 0.84 | 0.86 | 401 |
| Insights on Life | 0.74 | 0.73 | 0.74 | 224 |
| avg/total | 0.83 | 0.83 | 0.83 | 2115 |

Table 6. confusion matrix

| | | | | | | |
|---|---|---|---|---|---|---|
| 109 | 0 | 2 | 0 | 14 | 1 | 6 |
| 2 | 58 | 2 | 0 | 10 | 2 | 2 |
| 7 | 3 | 253 | 3 | 31 | 4 | 7 |
| 3 | 0 | 3 | 37 | 10 | 1 | 1 |
| 20 | 3 | 45 | 5 | 795 | 29 | 22 |
| 2 | 6 | 3 | 4 | 30 | 338 | 18 |
| 4 | 5 | 8 | 1 | 35 | 8 | 163 |

be concerned with these two elements: the first one is the sample data set is too small, and the second one is the data set labeling results are not good. We only use a single classification to label but the same paragraph may involve multiple aspects. If we use a 7-dimensional vector (each dimension uses 0 and 1 to indicate the presence or absence of the classification) instead of a single classification, we may get better results.

### 3.2. Text sentiment detection Based on Baidu AI open platform

In social networks, the emotions contained in texts published by users can greatly affect other users. Some users who always publish positive and optimistic trends will create a positive and optimistic personal impression in the

Table 7. python code

```
from aip import AipNlp
import json
client = AipNlp(APP_ID, API_KEY, SECRET_KEY)
text = 'this is test'
result = self.client.sentimentClassify(text)
sentiment = result['items'][0]['sentiment']
```

minds of people, and vice versa. This subsection will classify the emotions of the text to explore the impact of emotion on the heat.

Baidu AI open platform is a free online AI service launched by Baidu, which can realize natural language processing and image processing functions simply and quickly by downloading its SDK. We use the emotional tendency analysis to detect the emotion in the collected text data and classify the emotional type, and the result includes three types: the position type, the negative type and the neutral type. The codes of python for detect emotion using the SDK provided by Baidu AI open platform are shown in the Table 7 The APP_ID, API_KEY and SECRET_KEY here are needed to applied in Baidu Developers Center.

### 3.3. Image quality and aesthetic rating Based on Google NIMA model

As an important medium in social networks, images are often richer in information than words. A photo with exquisite composition, wonderful colors and superior quality is more attractive than a thousand words. In this sebsection, we try to quantify the quality and aesthetics of images to explore the relationship between dynamic images and heat in social networks.

Quantification of image quality and aesthetics is a problem in image processing and computer vision. Image Quality Assessment processes the pixel-level degradation problems like noise, blur and compression distortion. Aesthetic evaluation is the extraction of semantic level features related to emotion and beauty in images.

Some test photos from the large-scale database for Aesthetic Visual Analysis (AVA) dataset, as ranked by NIMA, are shown in the Figure 4. Each AVA photo is scored by an average of 200 people in response to photography contests. After training, the aesthetic ranking of these photos by NIMA closely matches the mean scores given by human raters. We find that NIMA performs equally well on other datasets, with predicted quality scores close to human ratings. This paper is based on the open source code on github and use the pretrained model based on the AVA data set, and implement photo rating of the NIMA system on social networks.

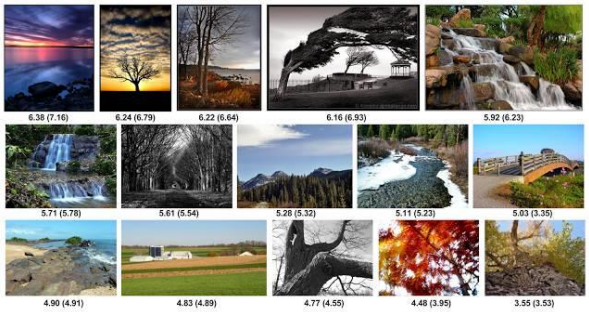


Figure 4. the AVA dataset

Table 8. time period division

| id | name | time period |
|---|---|---|
| 0 | midnight | 00:00 04:00 |
| 1 | dawn | 04:00 08:00 |
| 2 | morning | 08:00 12:00 |
| 3 | afternoon | 12:00 16:00 |
| 4 | Evening | 16:00 20:00 |
| 5 | night | 20:00 24:00 |

### 3.4. Object Detection in Image

We use Faster R-CNN to detect objects in images. The model we used is trained on MSCOCO, which contains 91 categories of objects. We extract the objects detected in an image and use them as a feature of the "moment". To use the feature in the clustering, we represent the supercategory exists in the image as 1 and represent it as 0 otherwise.

### 3.5. Dynamic Time Classification Generation

For most media information that needs to be disseminated, the time when the information is sent is critical because it determines the size of the group that can read the information to some extent. This paper attempts to mine the dynamic time of the user to explore the impact of the time of publication dynamics on the dynamic heat. We divide a day into six time periods as Table 8:

### 3.6. Nonlinear Fitting Based on Xgboost

Through the above five steps, data such as text, image, time and the like in the social network are cleaned and converted into corresponding continuous values or discrete values. In order to explore the common influence of the above five indicators on heat, this paper uses xgboost-based integrated learning method to fit all data.

**Data Source** The 14462 data collected from the QQ zone, after a series of data cleaning, transformation and fusion, is finally filtered to 1982 relatively complete data for

Table 9. RMSE on training set and test set

| data set | RMSE |
|---|---|
| training set | 0.002 |
| test set | 1.997 |

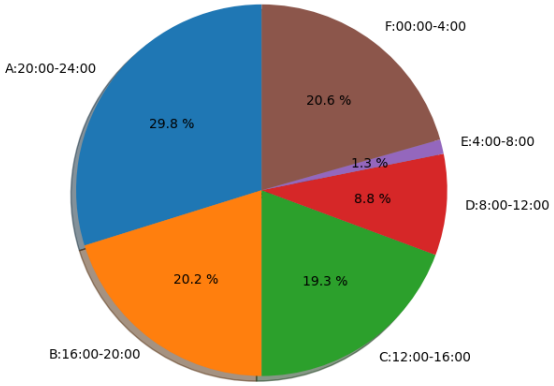


Figure 5. The user publishes dynamic time classification statistics and often stays up late

training. And the data is divided the training set and the test set according to the ratio of 7:3.

**Results** The results of the experiment were evaluated using the root mean square error (RMSE). The results are as shown in Table 9

**Result Analysis** The experimental results represent over-fitting, which is related to many factors, for example the data set is too small, the text classification is not accurate enough, the image classification is not accurate enough, etc. But the main factor is that it is not enough to understand and extract dynamic content. Due to the complexity of the language itself and the large number of ”expression packs” in the online community, information extraction becomes extremely difficult.

## 4. Construction of user portraits

### 4.1. User-based Image Construction Based on Statistics

In the previous sections, we describe the method of extracting five kinds of effective information from the ”moments” in detail. These extracted information can be used for the construction of user portraits after some statistics and analysis. For example, if the user publishes a dynamic time classification statistical result in the Figure 5, it can be found that the user often stays up late.

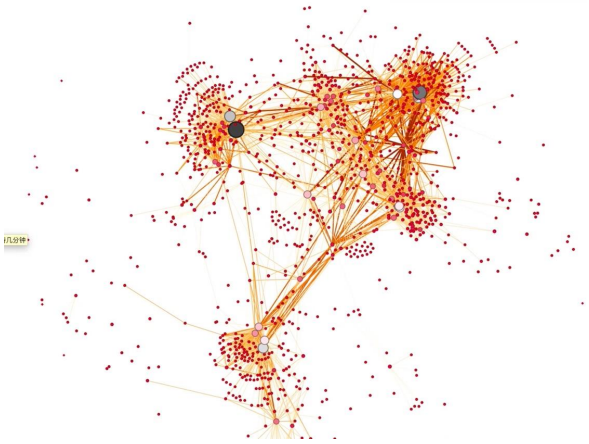

Figure 6. friends clustering

### 4.2. User Image Construction Based on Clustering

By modeling and clustering the friends and common groups between friends of the users, we got the following results which is shown in Figure 6

## 5. Conclusion

This paper mines and analyzes the data in acquaintance social networks through natural language processing and image processing. We use a variety of algorithms and models for deep learning and machine learning, involving data acquisition, data cleaning, transformation, fusion, mining and many other processes. However, due to limited time, the experimental results are not perfect, and it is difficult to draw valuable information. But it still puts forward a variety of ideas and methods for discussing ”what kind of content in acquaintances social networks are more popular with friends”. We hope to continue relevant study in the later period.