# 基于熟人社交网络的文本与图像数据挖掘

摘要：与微博、知乎、贴吧等公众的网络社区不同的是，朋友圈、QQ空间等网络社区是以相互熟识的朋友、亲戚、同学等为基础建立起来的，因此，我们称这类社区为“熟人社交网络”（与之相对的，称为陌生人社交网络）。在熟人社交网络中，由于用户之间除了线上的交流之外，往往还有线下的更为直接的接触，这使得熟人社交网络与陌生人社交网络中的交流方式变得不同。随着互联网的飞速发展，普通用户越来越需要在熟人社交网络中树立一个良好的印象，通过发送“动态”来引起好友们的关注。本文以QQ空间中采集得到的真实数据为基础，使用热度来衡量每一条动态的受欢迎程度，从而对熟人社交网络中如何产生更受欢迎的动态进行了建模和分析，并提出了一种基于用户在熟人社交网络中发表的文本、图像等数据构建用户画像的方法。

关键字：熟人社交网络、自然语言处理、图像处理、用户画像

1. Introduction

中文网络社区大致可以分为两类，一是以微博、知乎、贴吧为代表的陌生人社交网络，在这类社交网络中，每个用户发表的内容可以被绝大多数用户随意查看，不管用户彼此之间是否认识，因此，用户在此类社区中发表的内容通常是经过大量美化和修饰，真实性难以判断；二是以朋友圈、QQ空间为代表的熟人社交网络，这类社交网络中的用户发表的内容仅能被该用户的好友查看，这些好友通常有同学、同事、亲戚、朋友等关系，因此，此类社区中的内容更贴近用户真实的想法和性格。通过这些内容构建用户画像，能更好的区分用户群体，以达到精准营销的目的。另一方面，随着互联网的飞速发展，社交网络与人们的日常交流的联系更为紧密，普通用户也迫切的需要在熟人社交网络中营造一个良好的形象，通过发表一些经过修饰和美化的动态来树立一个良好的形象，然而与陌生人社交网络有很大差异的是，由于接收该动态的用户通常对该用户有现实中的接触和了解，所以在陌生人社交网络中能赢得大量热度的方式在熟人社交网络中并不奏效，比如在朋友圈中发表与真人完全不符的自拍更容易引起熟人们的反感。因此，本文通过机器学习等方法针对熟人社交网络中产生的内容与热度的关系进行建模，探索在熟人社交网络中发表什么样的内容更受欢迎。本文以热度来衡量受欢迎的程度，它是一种基于浏览量、点赞量与评论量的值，定义请参考第2.2节。本文以QQ空间中的动态为例，从以下几个指标对每一条动态进行量化和分析：

1. 动态中包含的文本的内容
2. 动态中包含的文本的情绪
3. 动态中包含的图片的质量与美感
4. 动态中包含的图片的内容
5. 用户发表该动态的时间

此外，这些指标也将用于自动构建每一位用户的用户画像。

下面将对如何提取和构建这些指标进行详细描述

1. 概念与定义
   1. 数据来源

从QQ空间中采集的从2014年6月10日至今的“说说”内容，所有被采集的用户均已清楚并同意本研究，累计采集的动态共14462条，记为D。经数据清洗之后，每条动态主要包含以下字段

|  |  |  |
| --- | --- | --- |
| 字段名词 | 解释 | 类型 |
| Tid | 每条动态的唯一标识 | string |
| **like\_num** | 点赞数量 | int |
| Prd\_num | 浏览量 | int |
| Cmt\_num | 评论量(不包括每条评论的回复) | int |
| Cmt\_total\_num | 总评论量(包括每条评论的回复) | int |
| Time\_stamp | 发表该动态的时间戳 | Int |
| content | 每条动态的文字内容 | string |
| Uin\_list | 该条动态的点赞的人的列表（包含昵称和性别） | json |
| Cmt\_list | 该条动态的评论的人列表 | json |

* 1. 热度(E)的定义

设D中的平均每条动态的评论量为m, 点赞量为n，浏览量为h，则评论量与点赞量的转化关系为a, 点赞量与浏览量的转化关系为b，其中，

a = n / m

b = h / n

则，对于每一条动态有熵e，

e = mi \* a \* b + ni \* b + hi

其中mi 、ni 、hi分别表示第i条动态的评论量、点赞量与浏览量，

则，每一条动态的热度Ei，定义熵e归一化后的值，即：

Ei = (ei – emin) / (emax - emin)

其中emax表示产生第i条动态的用户的最大熵, emin表示产生第i条动态的用户的最小熵。

因此，在假设每个用户的好友数量不变的情况下，E值的大小与好友数量无关。

1. 实现
   1. 基于RNN的文本内容分类

熟人社交网络中的文本类型偏向简短、零散，且经常带有语义不明的特殊符号，从而导致语义复杂多样。在本文中，我们自行标注和构建了一个包含7000多条文本的数据集，并使用递归神经网络（RNN）模型进行字符级别的训练和学习，最终F1值达到了0.83。

* + 1. 数据集的搭建

从原始数据集D中随机采样取出7230条文本数据，设立了7种文本类型，见下表：

|  |  |  |  |
| --- | --- | --- | --- |
| id | 类型名称 | 解释 | 例子 |
| 1 | 旅游与运动 | 包括外出旅行、户外运动或室内运动 | 真正的川藏线，从现在才开始 |
| 2 | 爱情与家庭 | 与爱情和家庭有关的任何事情，包括任何提到男女朋友、家人的事情 | 啊……以后就不要找我要我女票的照片了 |
| 3 | 学习与工作 | 与学习或工作有关的任何事情，包括任何提到学校、上课、公司、上班、同学、同事的事情 | 国际周让我深深的体会到什么叫“只要选课选得好……” |
| 4 | 广告 | 包括求点赞、求转发、卖东西或者各种微商的广告 | 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. | 生活日常 | 除了1/2/3/4之外的生活中的琐事，比如穿衣、饮食、住宿等 | 不如吃茶去 |
| 6. | 其他 | 文字过于简短或者没有文字；文本中不包含中文 | http://url.cn/WqvxQE |
| 7 | 人生感悟 | 关于人生或理想的言论或感概，通常不知所云；带有强烈感情色彩的文本 | 失落的岁月得到的美好垃圾的脖子. |

根据以上7种类型，对每条文本进行手动标注，最终得到的数据格式如下：

|  |  |
| --- | --- |
| 字段 | 解释 |
| label | 文本类型 |
| content | 文本内容 |

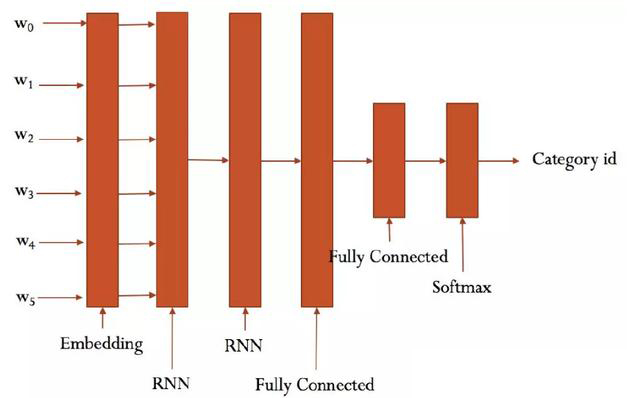
最终得到的各个类型的数据量如下：

|  |  |
| --- | --- |
| 类型 | 数量 |
| 1 | 440 |
| 2 | 229 |
| 3 | 1038 |
| 4 | 185 |
| 5 | 3051 |
| 6 | 1510 |
| 7 | 777 |
| 累计 | 7230 |

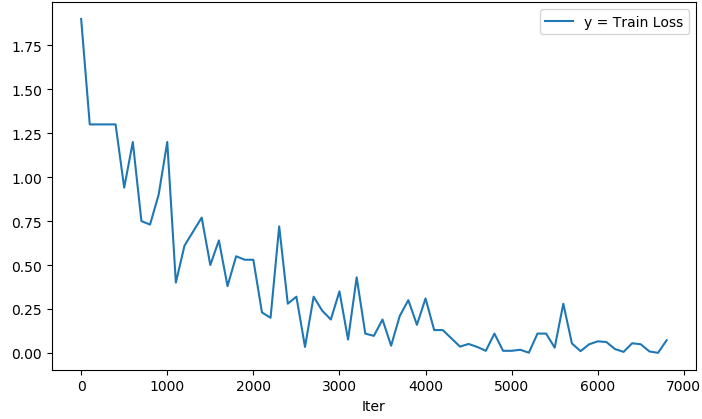
由于数据量比较少，所以按照7:3:3的比例划分训练集、验证集、测试集。

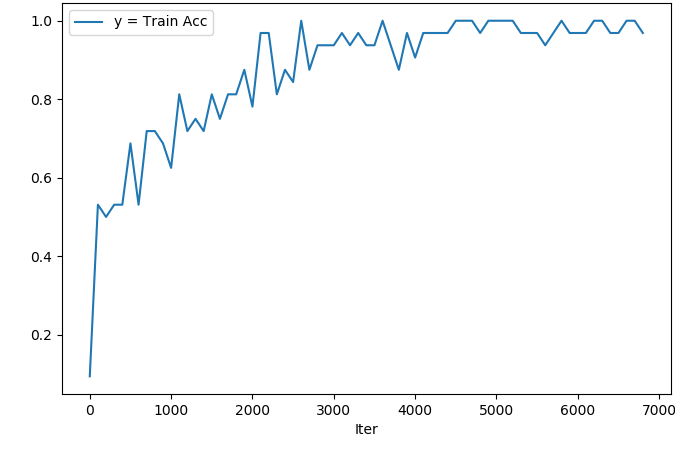
3.1.2 使用RNN进行训练

模型结构如下：



通过tensorflow编写的RNN神经网络，包含两层隐藏层，每层隐藏层包含128个神经元。损失函数与准确度（Accuracy）的收敛过程如下：





最终的测试结果如下：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 旅游与运动 | 0.74 | 0.83 | 0.78 | 132 |
| 爱情与家庭 | 0.77 | 0.76 | 0.77 | 76 |
| 学习与工作 | 0.8 | 0.82 | 0.81 | 308 |
| 广告 | 0.74 | 0.67 | 0.7 | 55 |
| 生活日常 | 0.86 | 0.87 | 0.86 | 919 |
| 其他 | 0.88 | 0.84 | 0.86 | 401 |
| 人生感悟 | 0.74 | 0.73 | 0.74 | 224 |
| avg / total | 0.83 | 0.83 | 0.83 | 2115 |

混淆矩阵如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

* + 1. 结果分析

最终的F1值效果不是很好，这可能与下面两个因素有关：

1. 样本数据集过少
2. 数据集标注结果不好，因为同一段话可能涉及多个方面，而我们这里只采用了单一的分类进行标注，如果使用一个7维的向量(每个维度用0和1来表示有无涉及该分类)来代替单一的分类，可能会取得更好的结果。
   1. 基于百度AI开放平台的文本情感检测

在社交网络中，用户发表的文本中包含的情绪会极大的影响其他用户，一些总是发表积极乐观的动态的用户会在人们心中树立一种积极乐观的个人印象，反之亦然。这一节将对文本的情感进行分类，从而探讨情感对于热度的影响。

百度AI开放平台是由于百度公司推出的免费在线AI服务，可以通过下载其SDK，简单快捷的实现自然语言处理、图像处理等功能，这里我们使用的是情感倾向分析，对采集到的文本数据进行情感检测，对其中包含的情感类型进行分类，分类的结果包括3种：积极、消极和中性。

使用百度AI平台提供的SDK进行情感检测的python代码如下：

**from** aip **import** AipNlp  
**import** json

client = AipNlp(APP\_ID, API\_KEY, SECRET\_KEY)

text = ‘这只是一段测试文本’

result = self.client.sentimentClassify(text)  
sentiment = result[**'items'**][0][**'sentiment'**]

其中APP\_ID, API\_KEY, SECRET\_KEY需要在百度开发者中心进行申请。

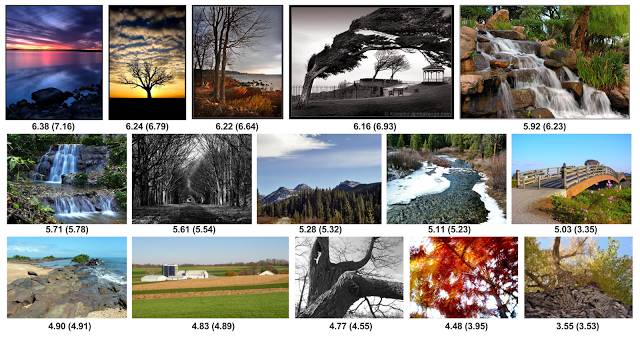
* 1. 基于Google NIMA模型的图像质量与美感评分

图像作为社交网络中的重要媒介，承载的信息量往往比文字更为丰富。一张构图精巧、色彩美妙、画质上乘的照片比千言万语更能吸引人的眼球，本节试图对图像的质量与美感进行量化，从而探索社交网络中动态包含的图像与热度的关系。

图像质量和美感的量化是图像处理和计算机视觉中的一个问题。图像质量评价（Image Quality Assessment,IQA）处理的是像素级的退化（degradation）问题，比如噪音、模糊、压缩失真等，美感评价为提取图像中与情感和美感相关的语义层次特征。

**参考：**[**https://blog.csdn.net/JtNbCOC8N2I9/article/details/78898510**](https://blog.csdn.net/JtNbCOC8N2I9/article/details/78898510)

Some test photos from the large-scale database for Aesthetic Visual Analysis (AVA) dataset, as ranked by NIMA, are shown below. 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.



本文基于github上的开源代码和基于AVA图片数据集的预训练模型，实现NIMA系统在社交网络中的照片评分。

* 1. 图像内容检测

。。。。。。

* 1. 产生动态的时间分类

对于大多数需要传播的媒体信息而言，发送该信息的时间都至关重要，因为这在一定程度上决定了能阅读到该信息的群体大小。本文试图对用户产生动态的时间进行挖掘，从而探索发表动态的时间对于动态的热度的影响。

本文对时间点进行划分，将一天中的时间分为以下6个时间段：

|  |  |  |
| --- | --- | --- |
| id | 名词 | 时间区间 |
| 0 | 午夜 | 00:00 – 04:00 |
| 1 | 凌晨 | 04:00 – 08:00 |
| 2 | 上午 | 08:00 – 12:00 |
| 3 | 下午 | 12:00 – 16:00 |
| 4 | 傍晚 | 16:00 – 20:00 |
| 5 | 晚上 | 20:00 – 24:00 |

* 1. 基于xgboost的非线形拟合

通过以上5个步骤，对社交网络中的文本、图像、时间等数据进行了数据清洗，并分别转化为对应的连续值或离散值。为了探究以上5个指标对热度的共同影响，本文使用基于xgboost的集成学习方法来对所有数据进行拟合。

* + 1. 数据来源

从QQ空间中真实采集得到的14462数据，经过一系列的数据清洗、转化、融合，最终筛选了1982条比较完整的数据进行训练，按照7:3的比例划分训练集与测试集。

* + 1. 实验结果

使用均方根误差（RMSE）对实验结果进行评估，结果如下：

|  |  |
| --- | --- |
| 数据集 | RMSE |
| 训练集 | 0.002 |
| 测试集 | 1.997 |

* + 1. 结果分析

实验结果出现了较大的过拟合现象，这与多方面的因素有关，比如数据集过小、文本分类不够准确、图像分类不够精确等。但主要的因素还是在于对动态的内容理解和提取不足够，由于语言本身的复杂性以及网络社区中大量存在的“表情包”等图片，这使得信息抽取变得极为困难。

1. 用户画像的构建

4.1 基于统计的用户画像构建

在前面的几个小节中，我们详细描述了从动态中提取5种有效信息的方法，这些经过提取的信息，再经过一定的统计和分析，可以用于用户画像的构建。比如对下图中用户发表动态的时间分类统计结果，可以发现该用户经常熬夜

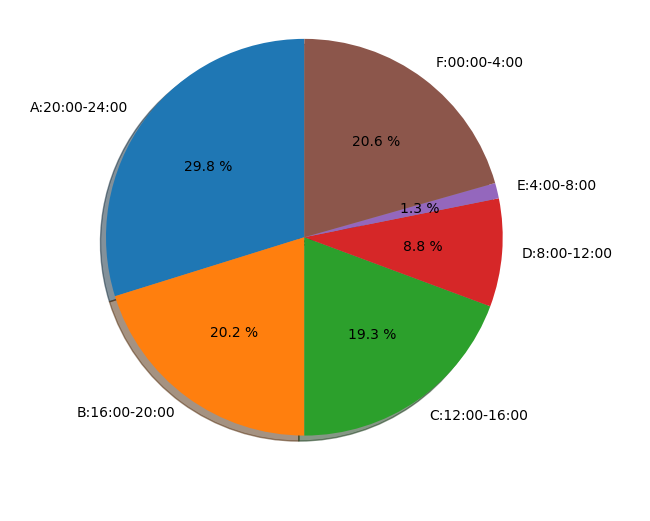


图. 该用户发表动态的时间分类统计结果常熬夜

4.2 基于聚类的用户画像构建

通过对用户的好友以及好友之间的共同群组的建模和聚类，我们得到了如下的结果：



该图的每个节点表示用户的好友和群组，每条边由群组指向好友，相近的节点具有更多相同的属性。通过对该图的分析，我们可以推断出用户属于的社交圈子。

1. 总结

本文通过自然语言处理、图像处理等方法对熟人社交网络中的数据进行挖掘和分析，运用了多种深度学习和机器学习的算法和模型，涉及数据获取、数据清洗、转化、融合、挖掘等诸多过程，但是由于时间有限，实验结果不够完美，难以得出有价值的信息，但是仍然为探讨“熟人社交网络中什么样的内容更受好友欢迎”提出了多种思路和方法，希望后期能继续进行相关研究。