数据挖掘大作业中期报告

——选题《社交网络中的个性化推荐系统》

宁小东 2120151024
黄建峰 2120150994
王新灵 2120151042

1. 问题描述

抓取微博中的用户属性、SNS 社交关系、过去 30 天内的历史 item 推荐记录等，预测接下来最有可能被用户接受的推荐 item 列表。

整个预测过程中涉及的各名词定义如下：

Item：是指微博中的特定用户，可能是个人、组织或集体，用于推荐给其他用户。例如，名人或知名组织可能会作为备选的“item 集合”推荐给用户；

推送（tweet）：是指用户将信息上传到微博系统的行为，或者是指推文本身；

评论：用户可以对推送（tweet）进行评论。评论不会像推送或分享（retweeting）那样显示给他的关注着，而是出现在该推送的评论历史中；

关注者（Follower）/被关注者（Followee）：如果用户 B 关注了用户 A，那么 B 就是 A 的关注者，A 是 B 的被关注者。

我们设计的数据集合包含如下字段：

i） 用户属性：包含用户 id、出生年份、性别、推送数、标签 id；

ii） Item：item id、item 种类、item 关键词；

iii） 用户行为（指分享或艾特等行为）：用户 id、行为目标用户 id、行为数、分享数、评论数；

iv） 用户 SNS 社交关系：关注者用户 id、被关注者用户 id；

v） 用户搜索关键词：用户 id、关键词。

利用以上用户信息，计算每个用户可能接受的推荐 item。

2. 数据抓取

使用爬虫软件在微博上抓取了用于训练和测试的数据，为了方便上传与计算起见，我们最终从 4.08GB 数据中选用了其中少量数据（使用 data_extract 函数，函数实现参见项目文件夹下的代码），包含 30000 条训练数据和 10000 条测试数据，选取的标准采用随机抽样的方式。

2.1 数据格式

使用如下格式抓取数据，数据共分为 7 大部分：

a） 训练集，格式为：（用户 id）\t（item id）\t（结果），其中结果的取值范围是{1, -1}，1 意为用户接受推荐，-1 意为用户拒绝推荐；

b） 测试集，格式与训练集相同，而（结果）的取值置为 0；

c） 用户个人信息，格式为：（用户 id）\t（出生年月）\t（性别）\t（推送数），出生年月为用户注册时填写，性别取值{0, 1, 2}，分别意为“未知”、“男”或“女”，推送数为整数，记录了用户推送消息的数量；

d） Item，格式为：（item id）\t（item 种类）\t（item 关键词），item 种类用“a. b. c. d”的格式写成，为分类的层级，关键词用字符串“id 1; id2; … id N”来表示；

e） 用户行为，格式为：（用户 id）\t（用户目标 id）\t（行为数）\t（分享数）\t（评论数）；

f） 用户 SNS 社交关系，格式为：（关注者 id）\t（被关注者 id）；

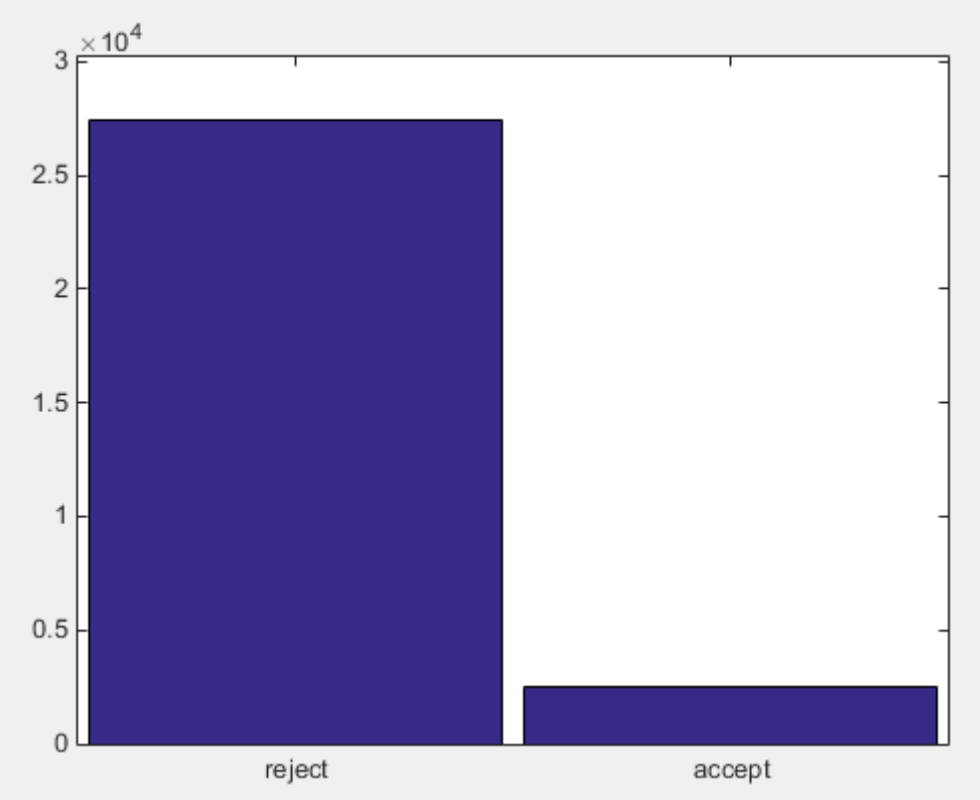
g） 用户关键词，格式为：（用户 id）\t（关键词）。

2.2 数据可视化分析

对逐个文件进行数据分析，分析的结果为：

（1）训练集

训练集包含用户 id，item id 和是否接受推荐等信息。首先观察接受和不接受推荐的频率：

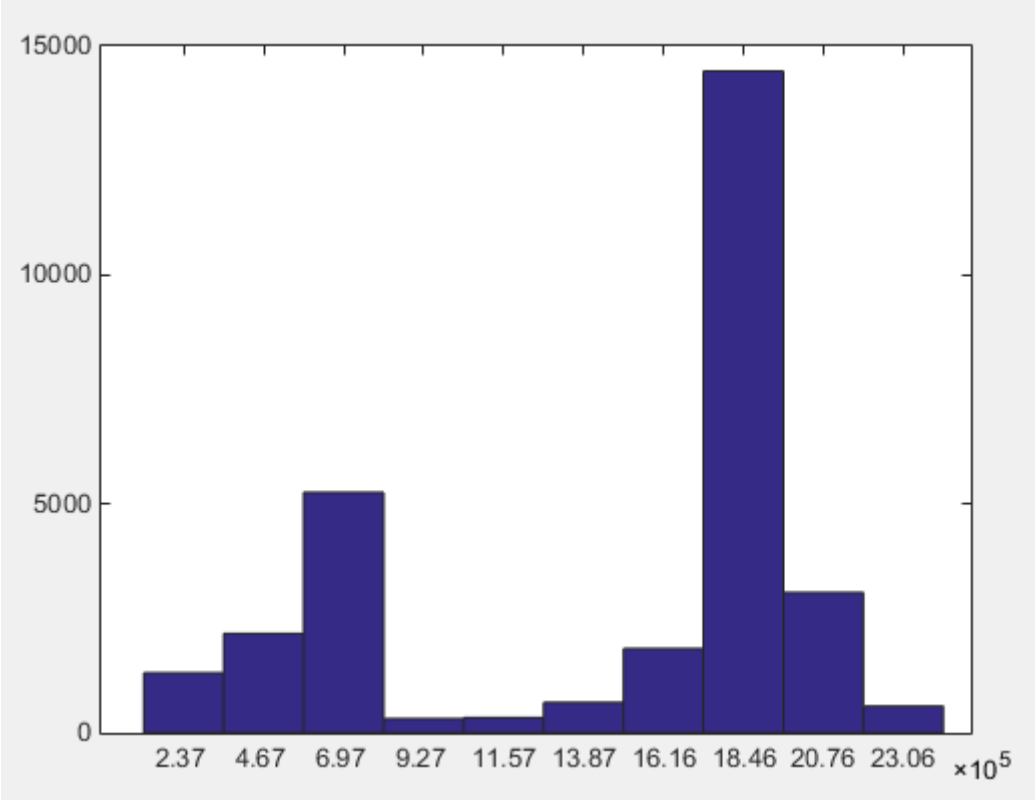


能够观察到推荐的 item 拒绝比例较高，其数据摘要为：

```
Show the acceptance in detail:
reject: 27456, 91.52%
accept: 2544, 8.48%
```

拒绝比例占到了 91.52%。

观察其 item 可得知：



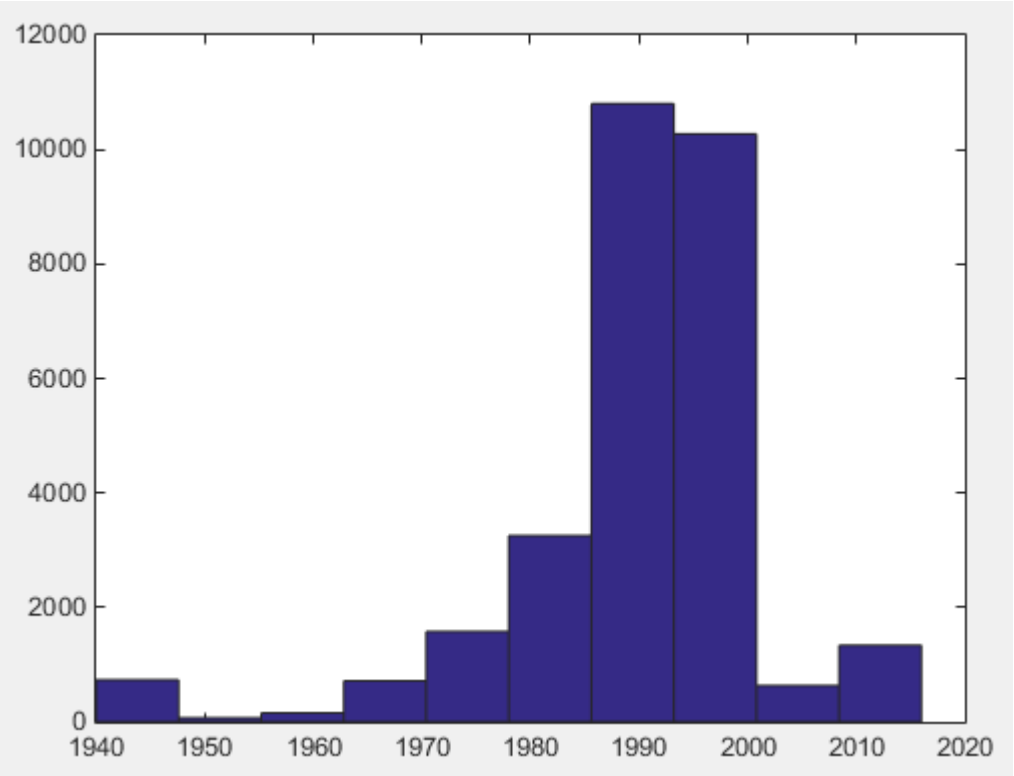
其 item id 取值在 237000~2306000 之间，获得其数据摘要：

```
Show the recommand item data abstract:
Most recommand frequency item id: 1775000
Most recommand item id: 1775000
    686 items in total.
```

在备选的 686 个 item 中，推荐次数最多的 item 为 1775000，且它也为用户接受频数最多的 item。从以上数据能够看出，**整体负样本较多，item 间推荐频次差异较大，在训练中可以减少推荐次数过少的 item 权重（例如 id 在 927000~1157000 的 item）。**

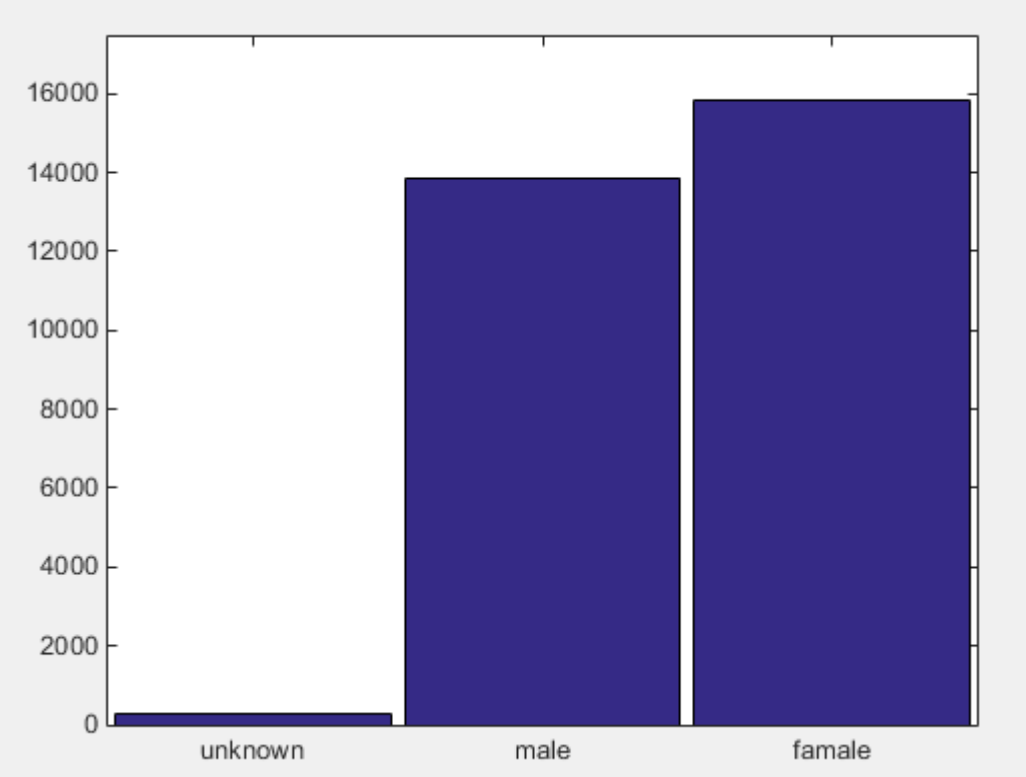
（2）用户个人信息

首先对用户出生年月进行可视化得：

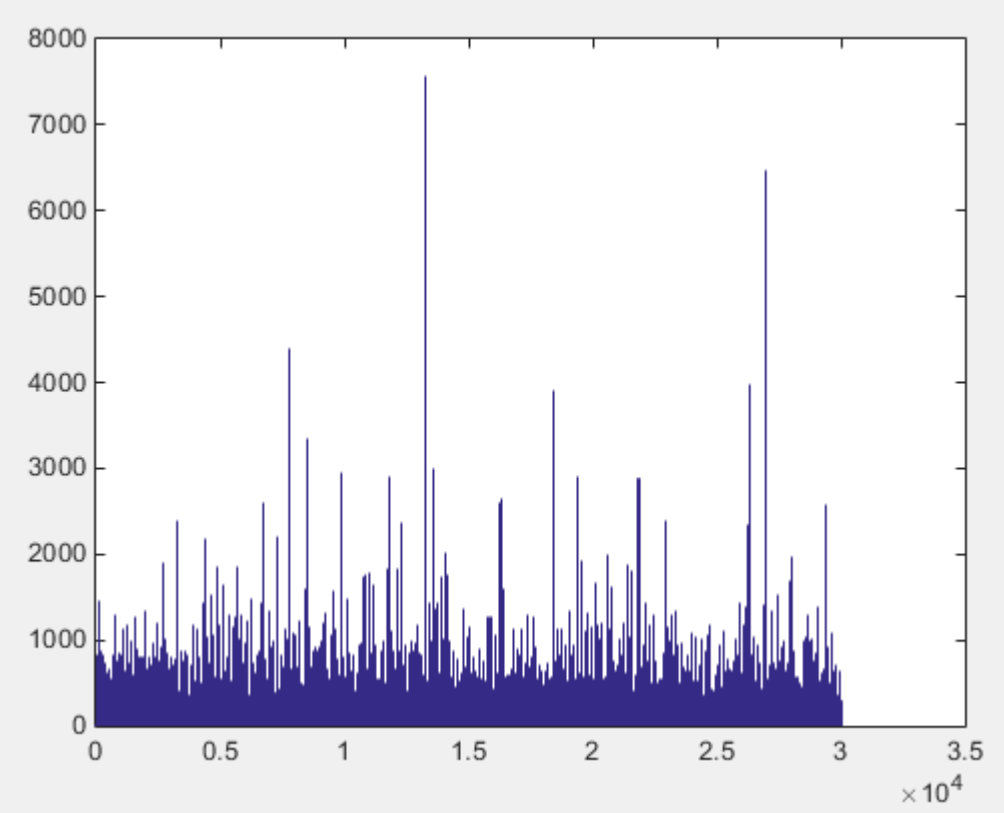


排除无效数据的干扰，我们不妨假设注册所填年月 1940~2016 为真实数据。获得出生年月分布如上图所示。

统计性别信息为：



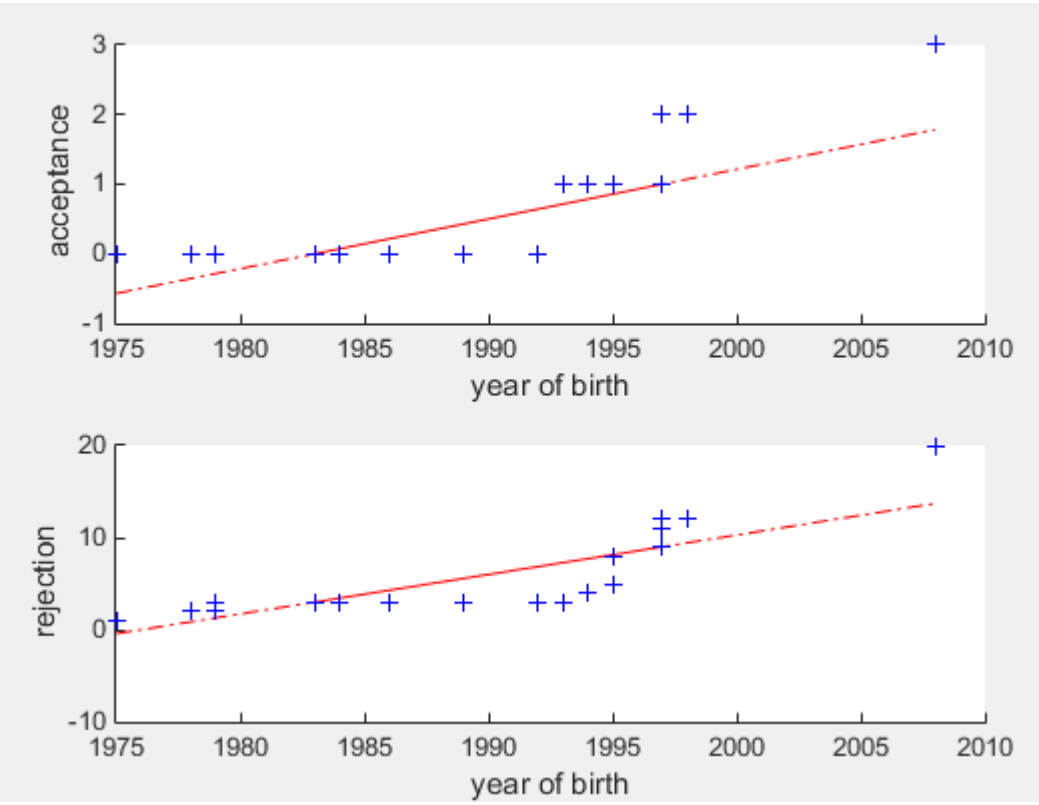
以及用户的推送总条数：

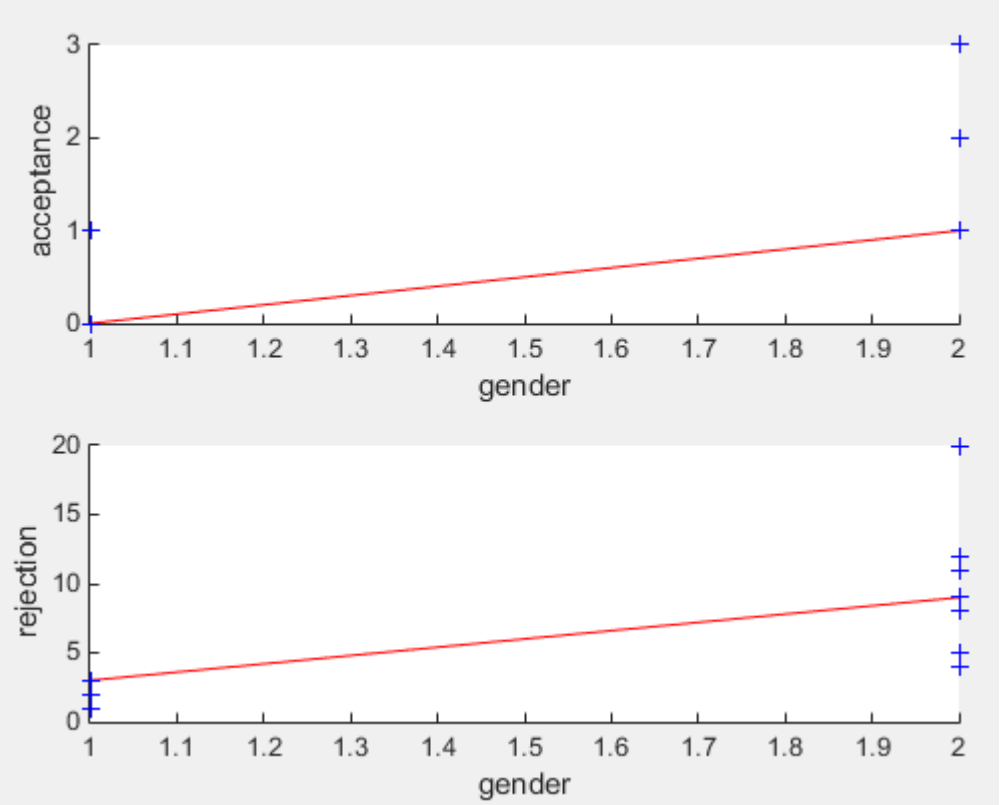


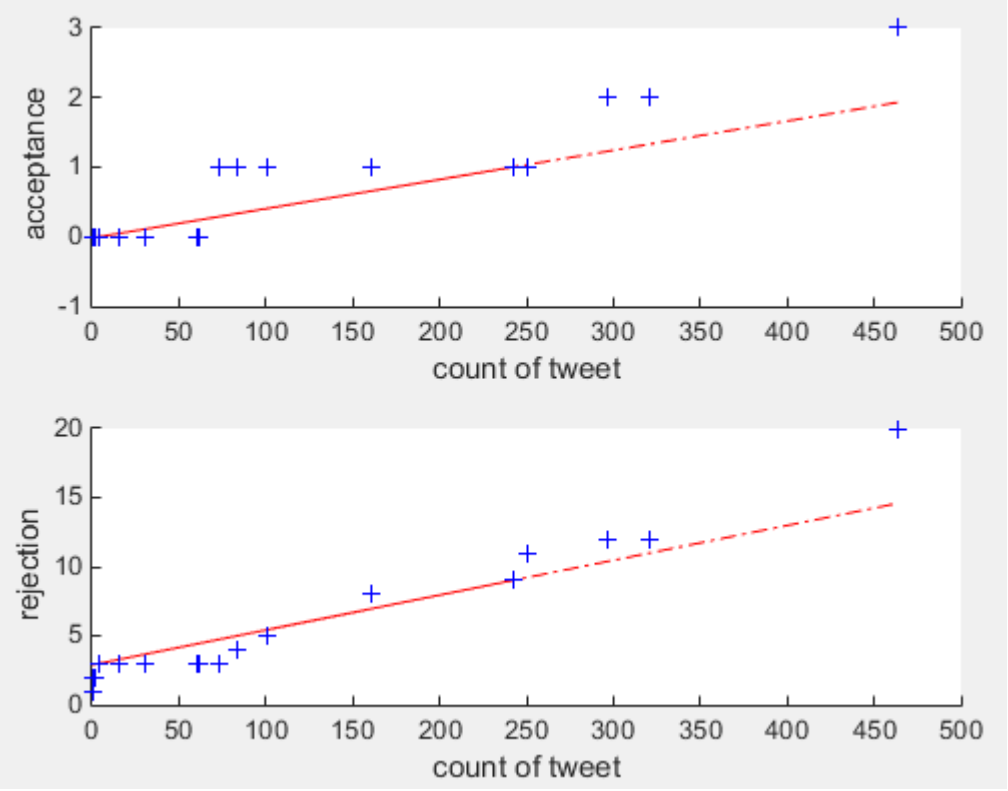
第 1~30000 个用户的推送个数差异较大。其中推送条数最多的 id 为：

```
The user who tweeted most frequent: 233114
```

将以上数据维度综合考虑，选择与用户“接受/拒绝推荐”最相关的变量。以抽取的部分用户为例，其中接受/拒绝程度分别与出生年月、性别、推送数的 QQ 图分别为：







分别计算它们与拒绝数(由于样本偏负,与拒绝的相关性更好则预测效果更好)的相关系数:

```
The coeffience between rejection and birth year: -0.02509
The coeffience between rejection and gender: 0.44451
The coeffience between rejection and tweet: -0.31414
```

能够看出预测过程中**最有代表性的因素为推送条数（相比出生年月，推送条数与拒绝数的负相关性要大的多），出生年月的影响较小；而性别虽然有较好的分辨性，但是样本偏度较大（取值为{0,1,2}），相对于复杂预测可能有不良影响，因此暂对性别因素不加以考虑**。

（3）item

Item 文件包含 item id，item 关键词等。

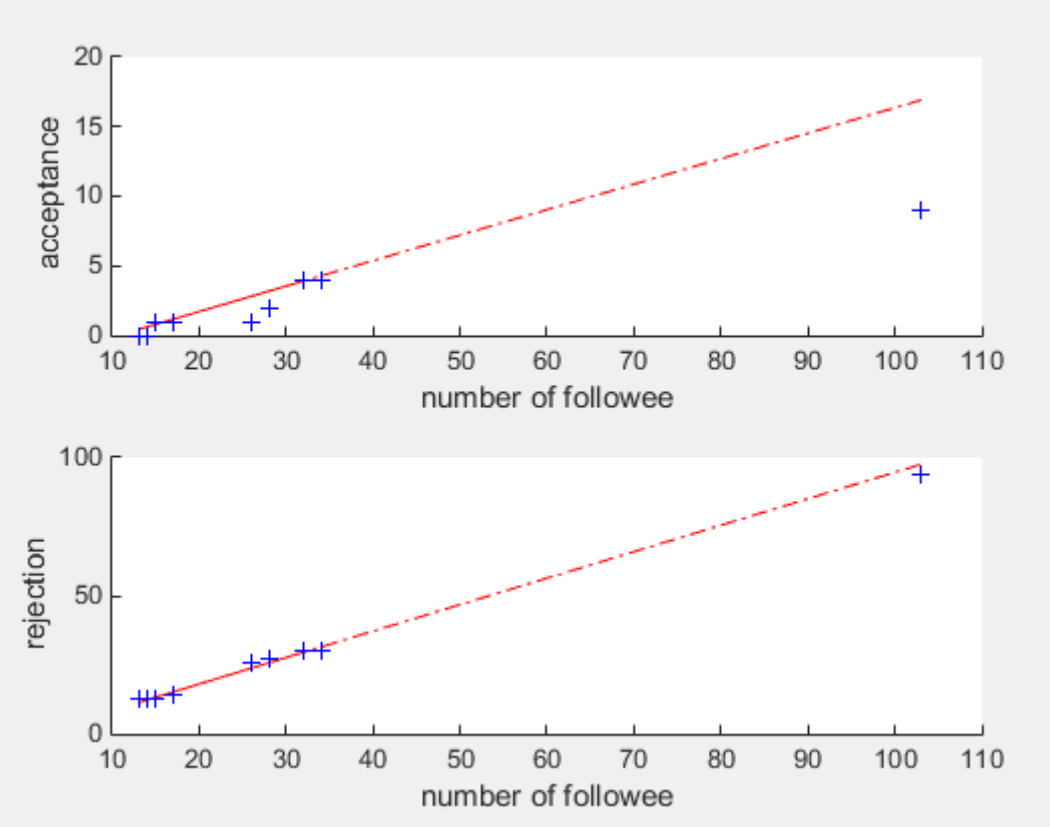
使用程序对 item 最热关键词进行获取可得：

```
Most frequent keywords: 3824
```

联系到用户感兴趣关键词，本部分数据可结合用户关键词进一步分析。

（4）用户 SNS

用户 SNS 包含每个用户的关注者/被关注者，计算用户所关注其他用户数量接受/拒绝推荐 QQ 图为：



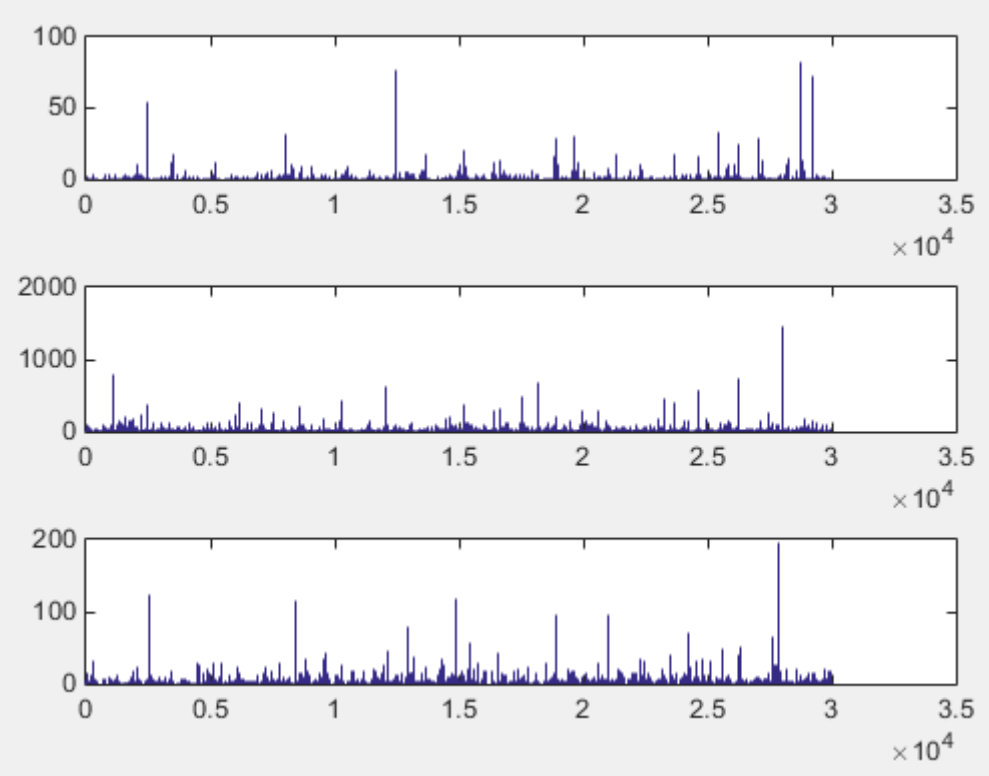
其相关系数为：

```
The coeffience between rejection and number of followee: -0.94489
```

能够看出关**注者数量拒绝程度有较高的负相关性，因此可以利用关注者数量对是否拒绝做出较好的预测**。

（5）用户行为

用户行为包含 3 中：艾特，分享或评论。对不同行为的统计如下图所示：



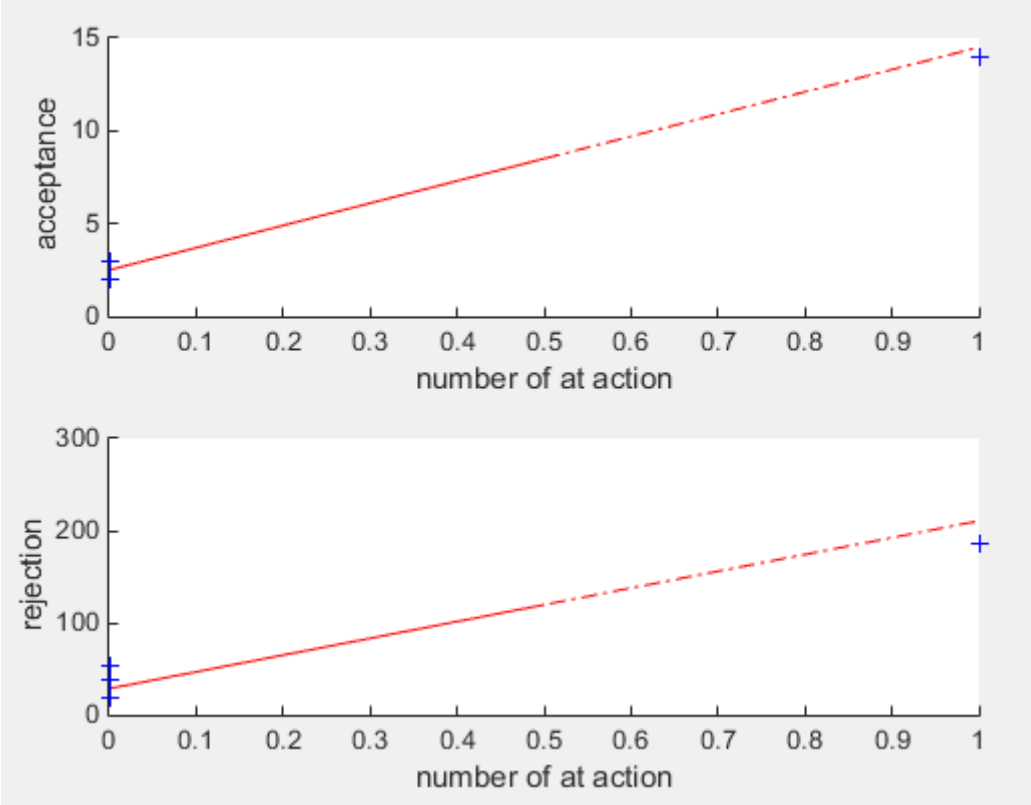
能够看出不同行为用户数值大小各不相同，每种行为的行为次数最多用户为：

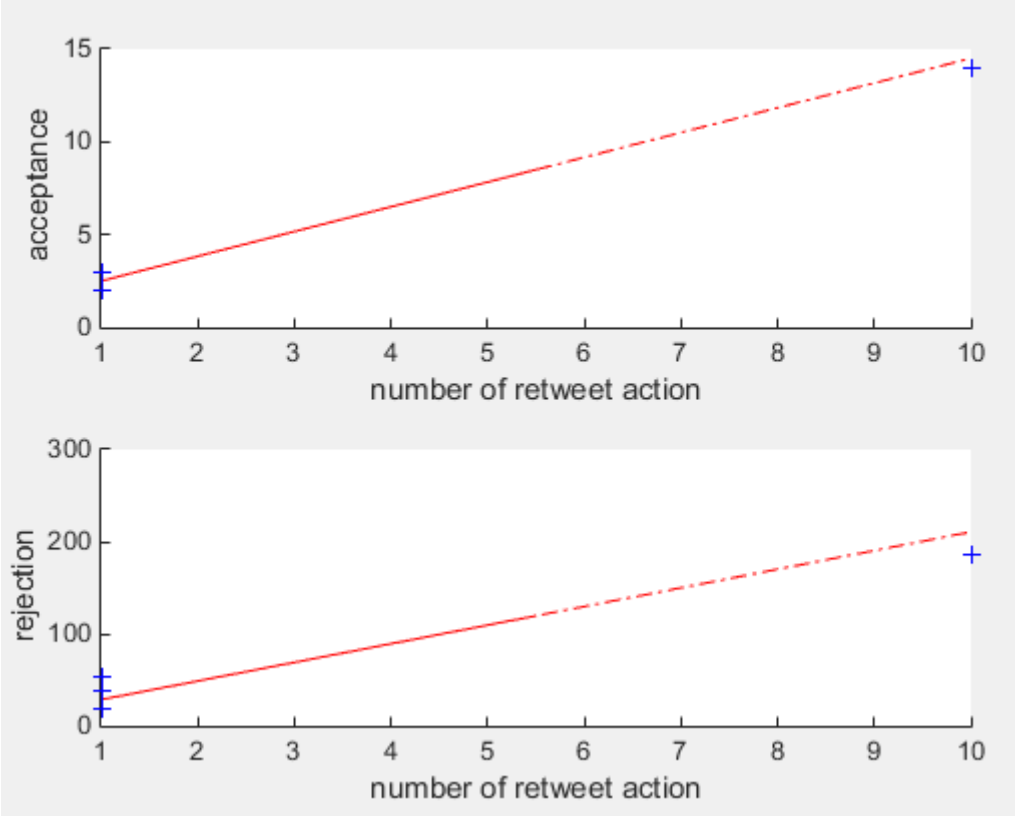
```
The user who at most frequent: 1004800
The user who retweeted most frequent: 1004700
The user who commented most frequent: 1004700
```

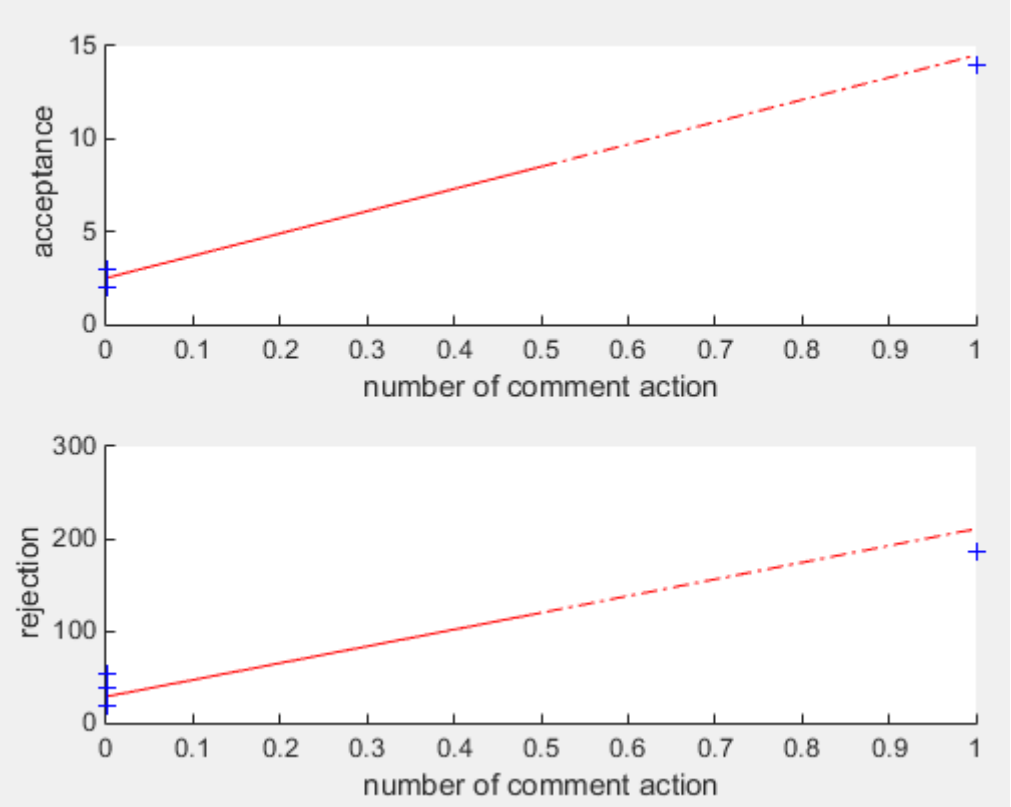
分享与评论最多者为同一人，而艾特数最多的并非同一人，三种行为的最高者并不重合。

对以上数据进行分析可知，不同用户三种行为的比重不同，可具体分析每种行为对 item 推荐接受程度的关联性。

对艾特、分享和评论分别与接受/拒绝程度画 QQ 图，并计算其相关性为：


acceptance
number of at action
rejection
number of at action


acceptance
number of retweet action
rejection
number of retweet action



三种行为与拒绝程度的相关系数为：

```
The coeffience between rejection and number of at action: -0.7746
The coeffience between rejection and number of rewteet action: -0.31843
The coeffience between rejection and number of comment action: -0.7746
```

三者中艾特行为与评论行为具有较高相关性，而分享行为相关性稍低。可以认为用户行为整体与用户接受/拒绝 item 相关，故在数据预测中应充分利用用户行为信息，从具有较高区分性的方面进行训练/预测。

2.3 数据清洗

为了高效地预测用户推荐 item，需要对抓取的数据进行清洗，使其更适合预测运算。

考虑到用户行为、item 关键词和用户 SNS 三者对推荐结果的影响较大，初步计划使用 user_action，item 和 user_SNS 作为主要的训练数据。据此对原数据进行降维处理，得到 18 维的标准训练/测试数据集（每条数据包含 x0~x17 共 18 个字段）为：

字段名	数据类型	含义
X0	Int	用于标记数据条数的常数，不参与训练/测试的计算过程
X1	Int	用户推送条数
X2	Double	用户感兴趣关键词或 item 关键词的加权求和
X3	Int	用户感兴趣关键词或 item 关键词的总数
X4	Double	用户所艾特的其他用户的加权求和（权重为来往行为次数）
X5	Double	用户所分享的其他用户的加权求和（权重为来往行为次数）
X6	Double	用户所评论的其他用户的加权求和（权重为来往行为次数）
X7	Int	用户所艾特的其他用户的总数
X8	Int	用户所分享的其他用户的总数
X9	Int	用户所评论的其他用户的总数

X10	Int	用户所关注其他用户的总数
X11	Double	用户所艾特的其他用户（包含目标 item 的情况）的加权求和
X12	Double	用户所分享的其他用户（包含目标 item 的情况）的加权求和
X13	Double	用户所评论的其他用户（包含目标 item 的情况）的加权求和
X14	Int	用户所艾特的其他用户（包含目标 item 的情况）的总数
X15	Int	用户所分享的其他用户（包含目标 item 的情况）的总数
X16	Int	用户所评论的其他用户（包含目标 item 的情况）的总数
X17	Int	用户所关注其他用户（包含目标 item 的情况）的总数

使用以上 x0~x17 共 18 维数据进行训练及预测。

3. 预测算法设计及流程

计划使用逻辑回归的方式分析数据。在整合所有训练/预测数据集后，对数据集进行以下步骤：

算法 1：逻辑回归

Step1：构造预测函数 h
计算边界 $\theta^T h$，后遭函数 $h_\theta(x)$

Step2：构造损失函数 J
通过最大似然估计取得对数似然函数 l(θ)
则 J(θ) = -1/m * l(θ)，其中-1/m 为系数

Step3：使用梯度下降法求最小值，
获取式 $\delta/\delta_{\theta j}$ J(θ)迭代得出最终的 θ 即结果

4. 下一步计划

下一步我们计划完成以下内容：

- 使用现有数据进行合并，去除冗余信息；
- 完成逻辑回归算法，对测试集进行 item id 预测；
- 完成实验撰写总结报告。