陈翊民 13307110219

COMP110031.01

2017/5/29

MATLAB数据分析期末作业: 基于移动应用陌陌用户数据

摘要

陌陌是近些年比较热门的基于地理位置的社交类移动应用。用户可以通过应用发送自己的生活动态，公开的动态可以被所有人查看，并且评论。用户还可以使用类似于豆瓣应用的功能，收藏自己喜欢的音乐，书刊和电影，并且根据爱好来结交朋友。但是和其他移动应用一样，陌陌的用户中包含大量垃圾用户，这些用户的数据不真实，对于研究用户行为形成了干扰，本文选择使用MATLAB2016b软件作为分析工具，使用包括机器学习中聚类算法在内的一些手段对数据进行清洗，提高数据的质量，方便之后的分析。这种方法也可以用于软件公司清理垃圾用户，提高平台质量，也可以帮助应用使用者鉴别一个用户是否为垃圾用户，以减少使用软件带来的潜在风险。

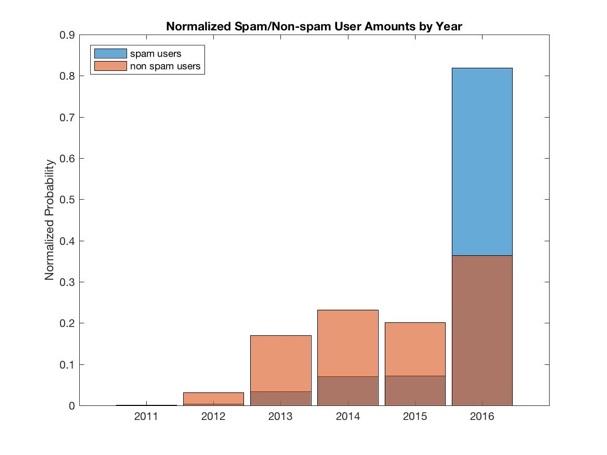
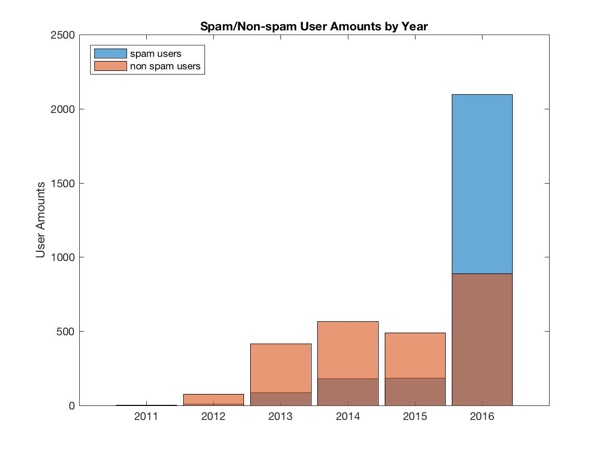
关键词： MATLAB Classification Learner App 机器学习 恶意用户 决策树

1. 问题背景

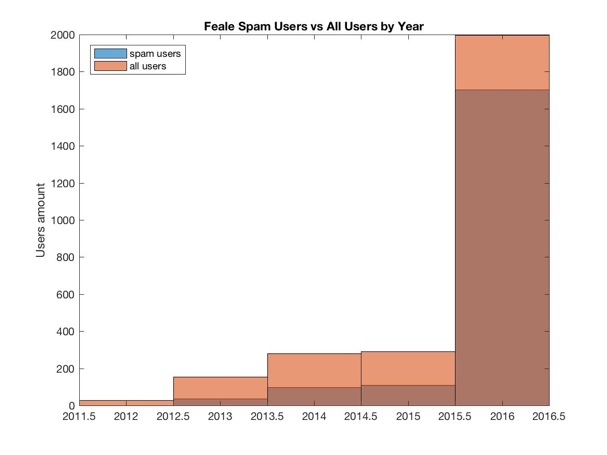
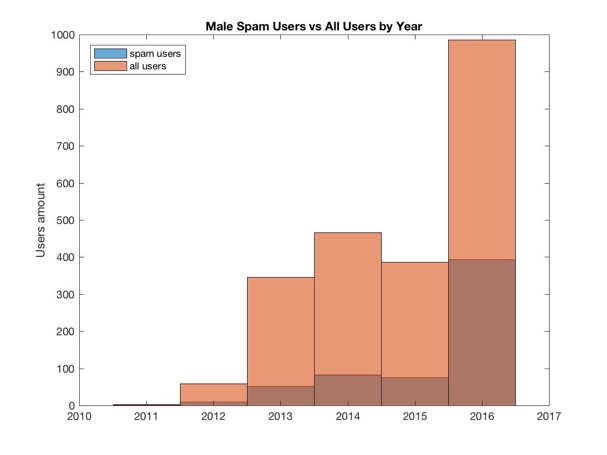
随着互联网技术的不断发展，人与人之间沟通的手段也迅速发展着。从信件到电邮，到再到微信等移动应用，社交软件（SNS）不仅功能越来越丰富，其发展方向也愈发多元化。有的社交软件专注于工作沟通，有的专注于购物和电子交易，有的专注于兴趣交友，有的基于地理信息，有的专注于直播互动。陌陌是国内最早一批基于地理信息的交友软件，在问世之初便吸引了大量用户，在业内有较大的影响力。陌陌的用户信息相比微信较为公开，适合利用其数据对用户行为进行分析研究。但是随着平台的壮大，用户人数增多，不少包括微商在内的恶意用户大量涌入，将此平台作为营销场所，同时一些性服务业的工作者也使用此平台寻找潜在客户，进行非法交易。种种原因使得这个平台的质量急剧下降，不仅使用这越来越难以通过这个平台认识到想要的好友，进行大数据研究的科研人员如果不加删选地使用用户数据，也难以获得满意的研究结果。综合这些考虑因素，我们认为基于本课题中已有的数据来生成一套判别恶意用户的手段，来筛掉恶意用户，提高平台质量，无论是对于陌陌软件公司，还是陌陌用户，还是数据科学家等科研人员，都是非常有必要的。因此，我的作业第一部分集中研究了恶意用户和普通用户在行为和信息方面的一些区别，并通过机器学习的手段生成了一套有效判别恶意的方法。第二部分在第一部分的基础上提取有效的用户数据来进行用户行为的分析，得到了不同年龄，性别用户的一些日常使用行为的分析结果，对于一些基于此的研究或软件开发有一些帮助和启发。由于对于此软件并不熟悉，不知道用户的信息是否都是真实的，或者真实性有多高。基于自己及身边的人在其他平台（如Facebook，Instagram，微信，豆瓣等）使用情况，我们假设恶意用户的所有信息都是不真实的，非恶意用户的信息都是真实的。我们之后的研究分析都是给予这一假设，所得出的结论正确性也依赖于这一假设的正确性。但是从已知数据没办法验证这一假设的真实性，我们的结论可靠性如何只能从逻辑和认知的角度来判断。

1. 恶意用户的分布研究
   1. 恶意用户在不同注册年份中的分布

在中国互联网行业，一般随着时间增加，一个软件的用户规模会扩大。但随之而来的是当一个软件的用户规模达到一定程度的时候，这个软件平台会吸引其他不怀好意的用户来此平台寻找机会。例如知乎应用初期，平台问答质量比较高，但是随着时间增加，知乎用户人数增多，同时在互联网上形成了一定的影响，越来越多的微商和代购涌入这一平台，造成知乎社区充满着各种商业气息浓厚的回答，社区质量也在不断下降。基于同样考虑，我们考察了恶意用户随注册年份的变化。如图1所示可以看出随着时间增加，非恶意用户的数目稳定增加，增长速度大致较平缓。而恶意用户速度则在2016年集中爆发。超过八成的恶意用户都注册在2016年，在此之前，陌陌用户中每年注册用户中恶意用户少于1/3，用户是带着正当需求来到这一平台的。而在2016年之后，注册用户中恶意用户超过2/3。或许是因为不少人看到了这一平台迅速发展后带来的“商机”，今而创建了一大堆营销账号大批涌入这个平台，从而获取不当利益。所以在选择好友的时候，对于来自2016年之后的用户需要格外警惕，统计数据告诉你超过70%的可能性该用户为恶意用户。如果是女性用户，则这个概率会更高一些。通过计算高大85%以上的2016年注册的女性用户为恶意用户，而在男性用户中这一比例不到4成。具体变化比例如图2所示。由此看来，用户是否在2016年注册为可以作为一个判定用户是为恶性用户可能性的较好参照指标。

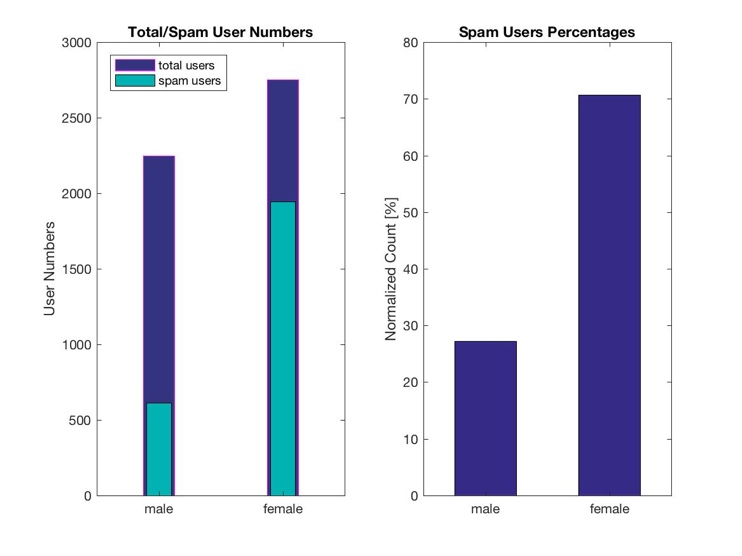


（图1 恶意用户与非恶意用户数目及比例随时间变化）



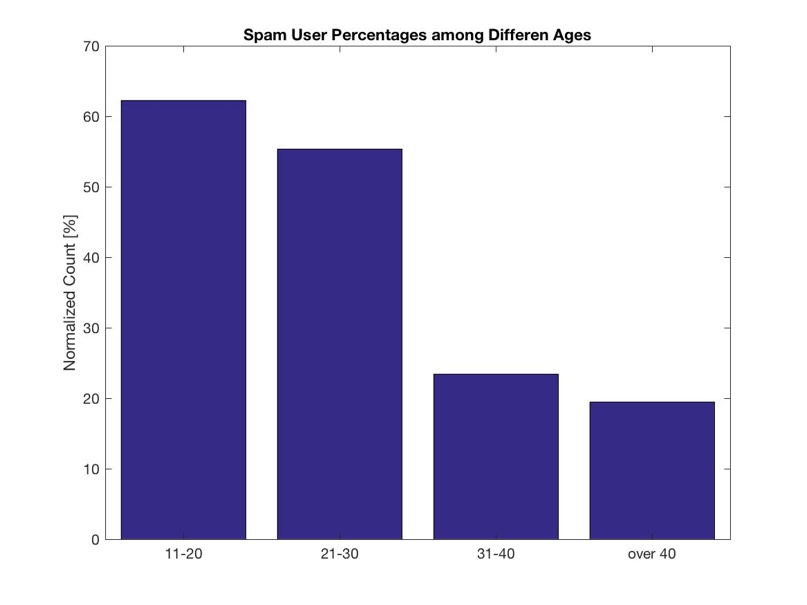
（图2 男性和女性用户中恶意用户人数随时间变化）

* 1. 恶意用户在不同性别中的分布

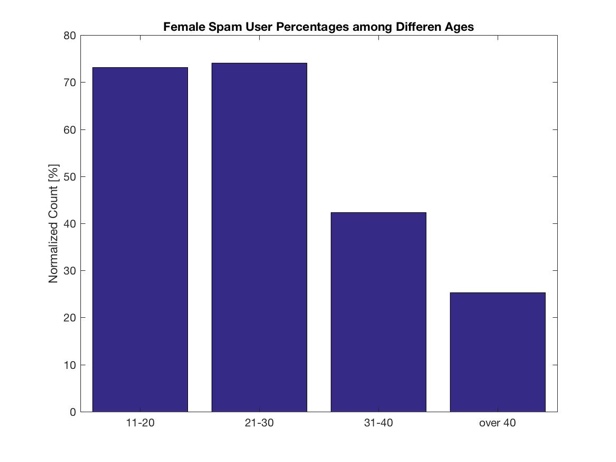
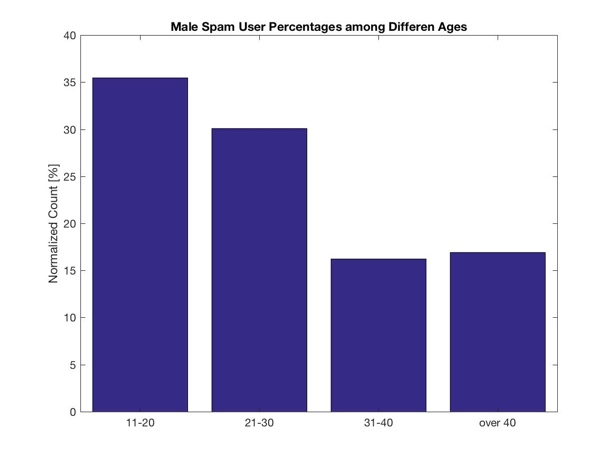
从恶意用户随时间分布的讨论中可以见恶意用户中性别显示为女性的比例高于男性，这个趋势在2016年变得愈发明显。在这一年中第一次男女性用户中第一次出现恶意用户人数超过非恶意用户。主要来源于2016年注册为女性的用户。综合统计各年份数据，我们得到图3，不同性别用户及恶意用户的分布图。可以看出恶意用户在女性中的比例远远高于男性中比例，这可能是因为在无论是真实社会还是在虚拟空间中，女性在社交方面相比男性更具有吸引力，从图3可以看出基于样本统计，女性用户收到的评论远超过男性用户。恶意用户通过注册信息为女性能吸引更多的人关注。从用户数量图中也可以看出非恶意用户中，男性数量远多于女性数量，恶意用户更有理由选择女性为公开性别。

（图3，恶意用户在男女不同性别中的分布）

* 1. 恶意用户在不同年龄段中的分布

通过分析数据，可以看出恶意用户的年龄段主要集中在11至30岁之间，因为恶意用户使用陌陌这一平台的主要目的并非是利用软件交友，而是通过软件吸引更多用户，进而获取经济及其他利益。伪装成年轻的用户能吸引更多其他用户，因而统计的数据从逻辑上来看是正确的。如果仔细分析用户数据，会发现总体恶意用户数据随年龄递减，但是具体到男女不同性别的数据时，两者会稍有差异。具体见图3。对于男性用户，恶意用户的分布与总体分布大致相同，而对于女性，有意思的是21-30岁的恶意用户比例相比10-20岁的恶意用户比例没有下降，反而上升了。个人推测这一年龄段段女性在社会上相比20岁之前的更成熟，感情经历更加丰富，对于男性用户来说吸引力更大，这或许是恶意用户选择自己年龄段在这一区间的原因。但是对于陌陌男性用户来说，这一年龄段女性中恶意用户比例高达3/4，这是非常不幸的高比例，你必须往更高的年龄段看才能找到更加靠谱的用户（作为恶意用户的可能性小于一半）

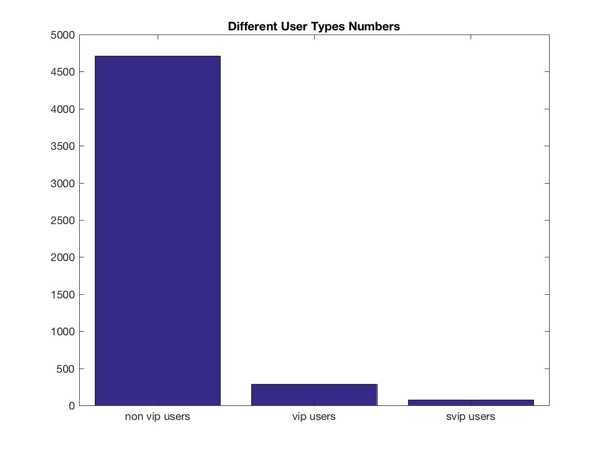
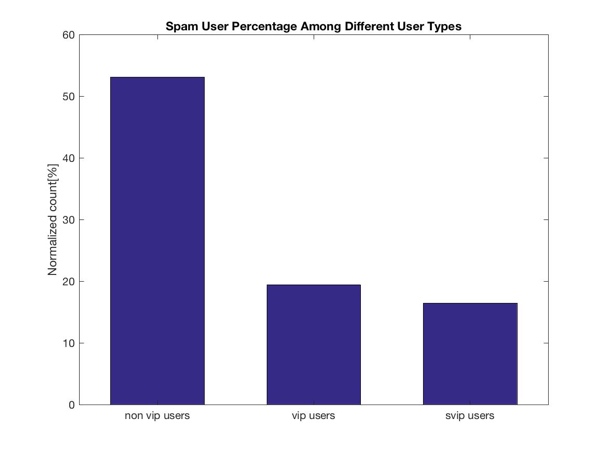
（图4，总体恶意用户在不同年龄组中的比例分布）



（图5， 不同性别恶意用户在不同年龄组中的比例分布）

* 1. 恶意用户在不同会员类型中的分布

用户是否愿意为提高使用体验而付费购买应用已经成为一个评价一个平台的优劣的指标。相比于普通用户，付费购买会员的用户更注重此平台的用户体验，因此付费用户中恶意用户比例更低。同时恶意用户并不关心平台用户体验，只希望吸引更多的其他用户来达到自己的目的。因而购买付费服务增加了成本，却并不能带来预期收益的更大提高。所以显然在付费用户中，恶意用户的比例应该比普通用户中要低，会员费越贵，会员中恶意用户比例会更低，这在数据中得到了体现。如图4所示。需要注意的，在整个平台中，付费用户数目极少，仅有5%以内的人付费购买了会员服务。所以如果只看绝对数据，普通用户中的非恶意用户数目其实是远超过会员用户的。所以如果仅仅凭一个用户是否是会员用户来决定是否与其成为好友，或者与Ta交流，其实会错过很多优质的非会员用户，而且说不定人家会员用户还看不上不是会员的你呢。所以，通过这种方法也不能很好地判断一个用户是否是恶意用户，最多只能作为一个判断其可能性大参考依据。



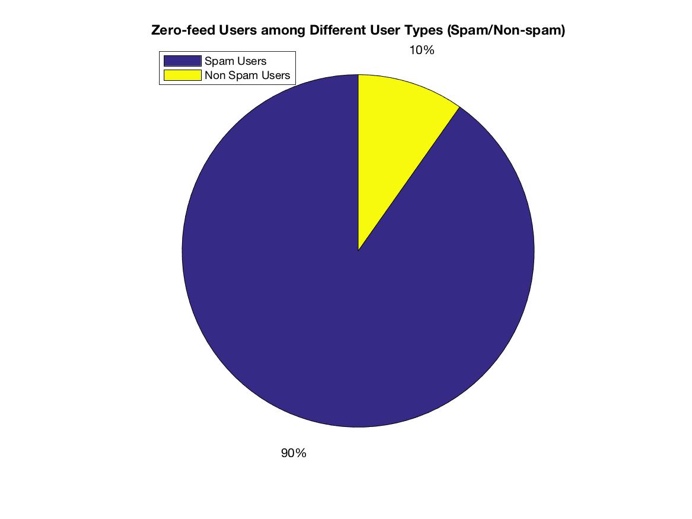
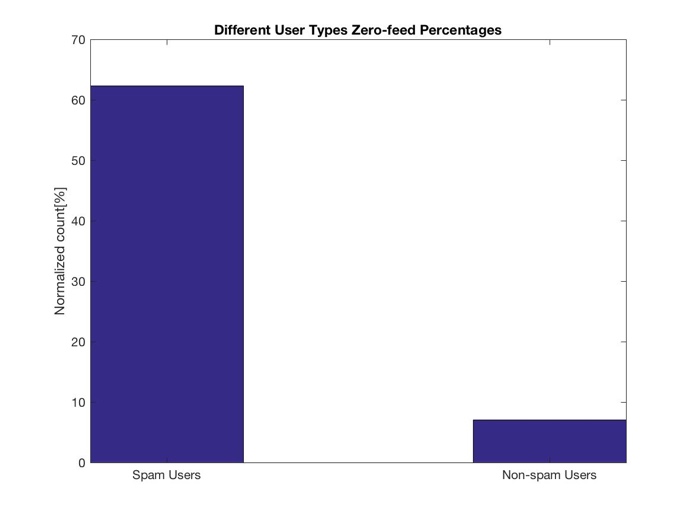
（图6， 不同会员类型中恶意用户的分布）

1. 恶意用户的行为研究

之前部分简单地分析了恶意用户在不同性别，不同年龄段，以及不同会员类型用户中的分布。虽然在10-30岁的非会员用户中女性恶意用户的比例较高，但是这部分用户基数也很大，这部分群体也是非常多的陌陌用户的目标好友群体。我们不能仅仅依靠年龄，性别，会员，注册年分等指标来筛选恶意用户，这也会造成较高的误杀比例。因而我们必须通过研究其他用户其他行为来进一步区别两者。

* 1. 恶意用户与普通用户的发帖比较

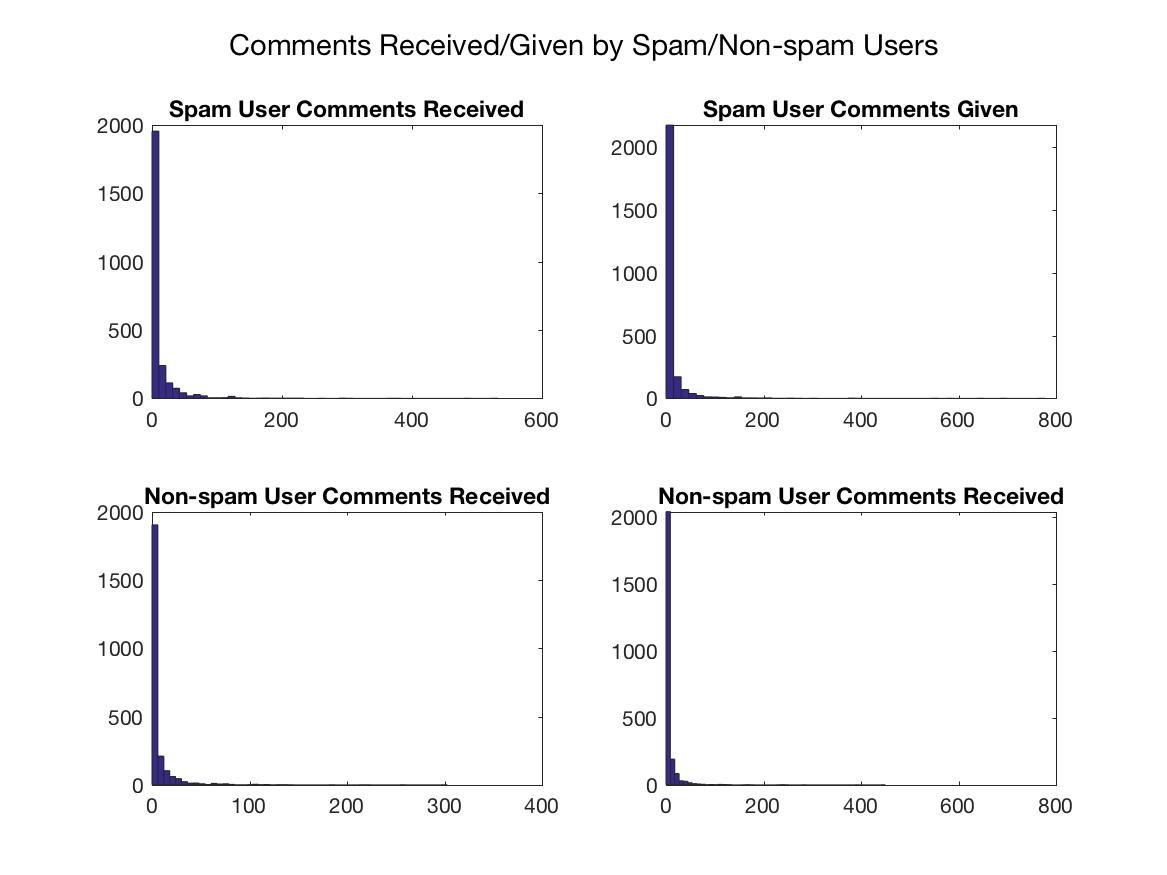
从统计数据可以看出，在恶意用户中，发帖数量为零的比例远远高于正常用户中。从发帖数量为零的用户统计分布来看，恶意用户的数目远远超过非恶意用户。在实际中，很多时候恶意用户是由一个群体大量生成，大量操作，因而与正常用户不同，没有大量时间来维护个人信息，分享个人生活动态。所以在实际使用中，当你遇到一个用户发帖数量为零的时候，你应该引起警觉，这个用户九成是垃圾用户，而且即便不是垃圾用户，你也无法从用户的分享状态中得知Ta的个人相关情况，在与其交流的时候需谨慎。所以，一个用户发帖数量是否为零也是一个判断该用户是否为恶意用户的好的参考指标。



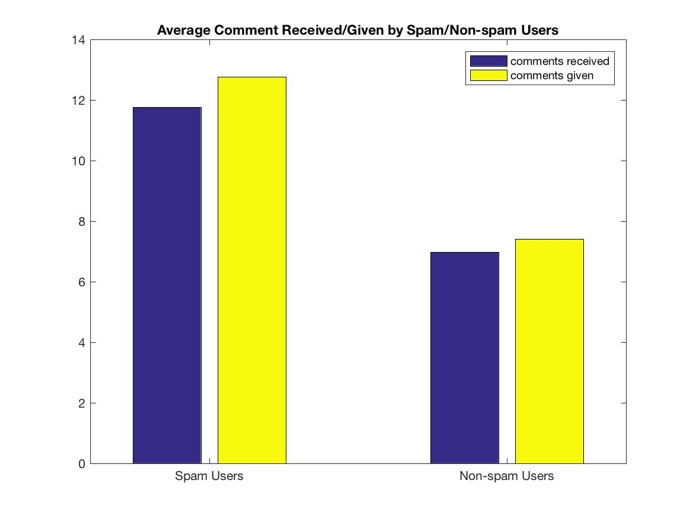
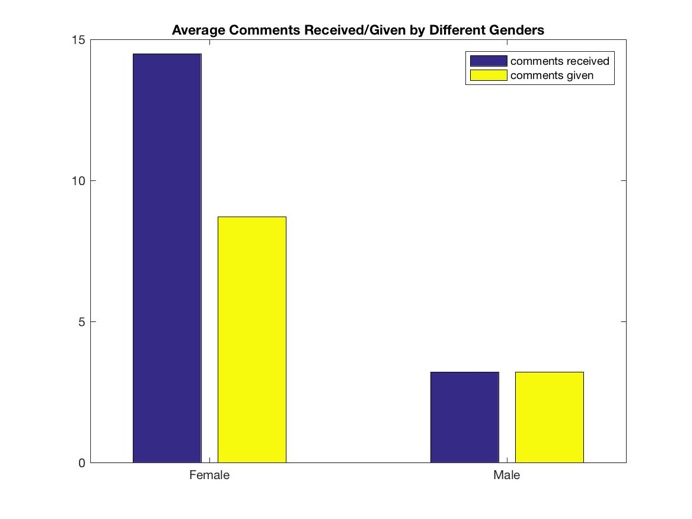
（图8，零动态用户中非垃圾用户与垃圾用户的数目）

* 1. 恶意用户与普通用户送出和收到评论数量的比较

从频率直方图来看，无论是垃圾用户还是非垃圾用户，收到和送出的评论分布大致相同，随着评论数目增加，人数递减。没有明显模式区别。

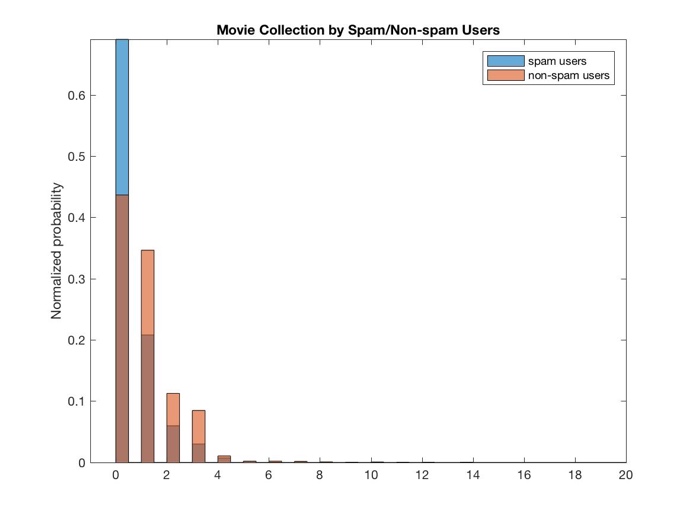


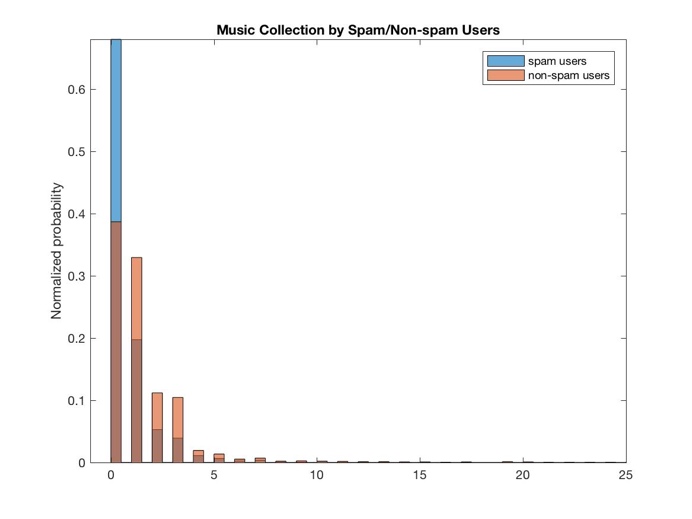
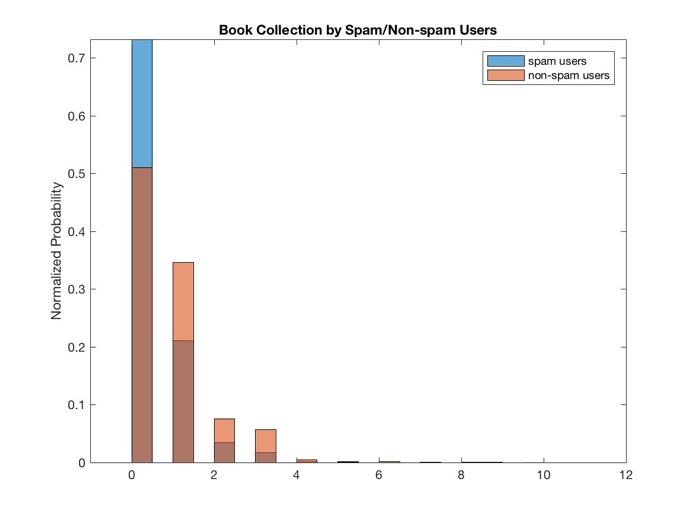
（图9， 恶意用户与非恶意用户收到和送出评论统计图）

针对此，我们通过统计用户平均收到的评论和送出评论的数目来试图区分恶意用户与正常用户。得到如下的结果。这表明无论是平均收到评论数量还是送出评论数量，恶意用户都远远多于正常用户，这似乎可以作为区分恶意用户和非恶意用户的一个指标。但是这存在两个问题，第一是这一指标只是平均值，只能说明平均意义下恶意用户比非恶意用户或送出的评论更多。我们无法选择一个合适的样本指标来作为判定依据来说明评论数目高于此数值的用户为恶意用户，低于此数值的为正常用户。此外统计还显示收到和发出的评论与用户性别具有很强的关联性。总的数据中女性用户收到或发出评论更多，女性用户中恶意用户的比例相比男性用户更高。说明评论树木统计意义下的差异是由性别引起的，还是由用户类型（恶意／非恶意）引起的。因此我们认为从评论数目并非是一个区别用户是恶意用户还是非恶意用户的好指标。

（图10，不同用户类型用户收到／送出评论（恶意／非恶意 男／女））

* 1. 恶意用户与非恶意用户兴趣爱好统计

我们使用给出的书籍，电影，音乐等阅读量来统计恶意用户与非恶意用户的差异。如图8所示。从图中可以看出，非恶意用户无论是在书籍，电影，还是音乐方面相较于恶意用户都有更多的收藏。从中反应非恶意用户希望通过兴趣爱好来与其他有着相同兴趣爱好的用户成为好友，所以会更加认真地维护个人兴趣方面的信息，但此方法对于短期内建立大量好友关系来说效率相对较低，因而恶意用户会通过其他手段来增加好友规模，不太在意此方法，所以数据显示恶意用户在各种收藏方面都明显少于非恶意用户。虽然统计数据显示恶意用户与非恶意用户在收藏爱好方面存在一些差异，但由于恶意用户与非恶意用户存在很大的重叠部分，因而这项指标作为判定一个用户是否是恶意用户也是不好的，准确性比较低。



（图11 恶意用户与非恶意用户在兴趣爱好方面的比较）

1. 综合分析及机器学习方法

前面的第二部分和第三部分，我们试图通过用户的注册年份，性别，年龄，是否为会员等身份特征，以及发帖数，评论互动，兴趣爱好等行为，结合我们的人生经验来综合分析判断一个用户是否为垃圾用户，所涉及到的指标非常多，判断起来非常复杂。我们只能在分析的基础上只能得到一个典型等垃圾用户的形象为，2016年注册的10-20岁非会员，发帖数为零的女性用户，对于一个任意给定的一组用户数据，我们用此方法不能准确地判断一个用户是否为恶意用户，特别是作为个人用户结交朋友，还是作为数据科学家分析用户行为，我们关注的是如何从一堆用户中较为准确地找出非恶意用户，至于能较为准确地判定一个用户是恶意用户其实实际意义并没有前者大。联系到现在大数据时代，我们经常通过一些机器学习的算法对挖掘到的数据进行一些清洗等预处理，我联想到是否可以通过机器学习到算法来达到这一目的。稍加分析便发现这其实是机器学习中最简单的分类算法，并且由于给定的数据中包含一个用户是否为恶意用户的信息，这个问题又是相对简单一些的监督学习（Supervised Learning），类似于此的问题很多。例如如何通过一定指标准确判断一个肿瘤是恶性还是良性，如何判断一封邮件是垃圾邮件还是垃圾邮件，等等。对于这类问题，有非常多大现成算法。而且MATLAB自带的工具箱十分强大，可以极大地降低机器学习的使用门槛。通过交互式界面选择数据和学习方法，然后进行学习和训练，最后可以通过工具箱输出此模型（函数／代码），直接用于下一次训练。

给定的数据样本为5000个用户资料，有两个用户的发出评论和收到评论数目异常高，为了达到较好的拟合效果，手动去除极少数异常样本。最后得到4998个样本，将其分为两组。前2449个用户资料为第一组，记为M1，用于学习和训练，产生预测一个用户是否为恶意用户的模型。后2449个用户资料作为第二组，记为M2，用每个用户资料的除去是否为恶意用户的这个参数的其余参数作为输入数值，再通过模型产生该用户是否为垃圾用户的结果，再将这一结果与用户是否为垃圾用户进行比较，通过比较模型的预测的准确率来判定所生成模型的可靠性，作为今后研究其他用户数据的可靠性参数。

在具体的操作中，我们使用MATLAB2016b中自带的Classification Learner工具箱。从momo.csv文件以table格式读取数据，然后在Classification Learner导入数据界面导入刚读取的数据，使用列为参数，选择spam这一栏参数为response，其他栏参数为predictor，使用5 folds Cross-Validation 防止过拟合。在模型选择的时候测试了各种Classifier，最后选择了在时间和效果方面均取得较好结果的Ensemble Classifiers中Bagged Trees进行学习。模型等Confusion Matrix如图12所示。95%的非恶意用户会被认为是非恶意用户，其中5%会被认为恶意用户。14%等恶意用户会被认为是非恶意用户，86%等恶意用户会被认为是恶意用户。在真实的数据分析中，被标记为非恶意用户中存在15%左右的恶意用户，非恶意用户中有5%的用户会被“误杀”。在一定程度上具有可行性。最后得到参数及结果如下。对于M2组数据，使用M1组数据训练出来的模型进行预测，准确率达到88.7%，这是相当高的水准，相比之前通过一系列指标进行判断的话。



（图12 训练产生模型等Confusion Matrix）

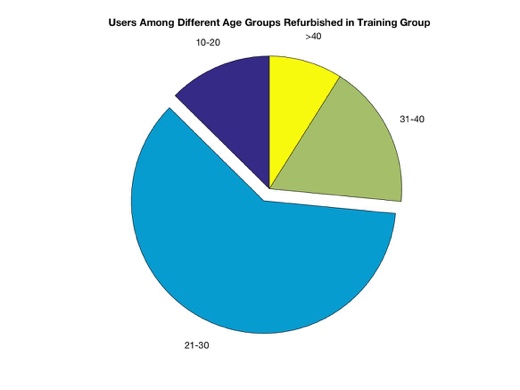
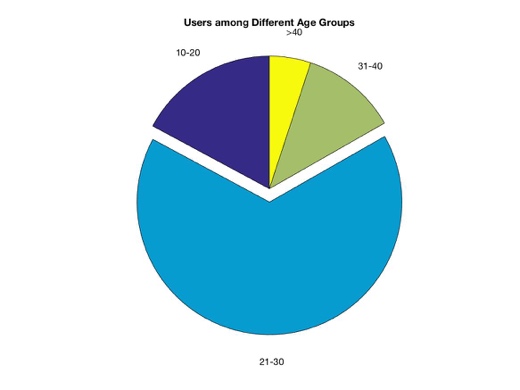
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Preset | Ensemble  Method | Learner  type | Number of  Learners | PCA | Accuracy | Prediction Speed | Training Time | Prediction Accuracy |
|  |  |  |  |  |  |  |  |  |
| Bagged Trees | Bag | Decision  Tree | 30 | Disabled | 90.1% | 7900 obs/sec | 14.725s | 88.7% |

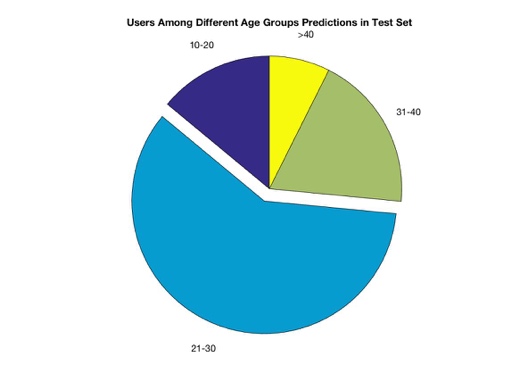
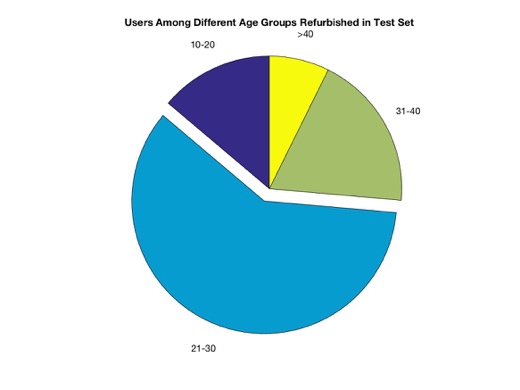
（表1 使用机器学习方法训练产生模型的参数与预测结果）

1. 非恶意用户使用行为分析及模型验证

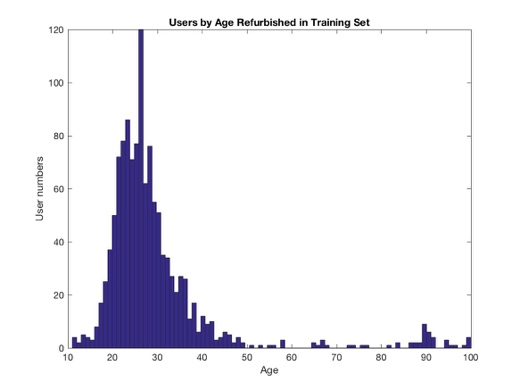
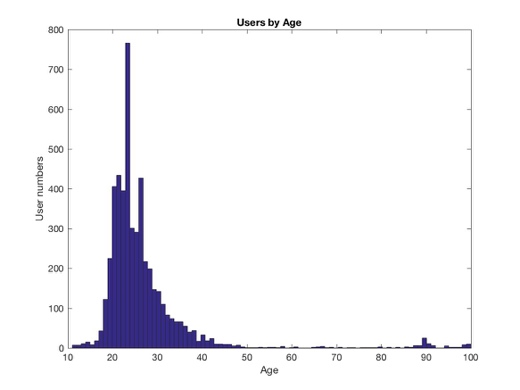
由于时间和精力有限，在最后一部分我们简单利用已有用户组中的非恶意用户数据对不同性别的行为进行了一些简单的分析，得出了男女用户在使用行为上的一些差异。同时运用产生模型预测了一批非恶意用户的数据，对此分析并与数据集中的非恶意用户数据进行了比较，说明了在大批量未标记的用户数据中，运用我们的模型产生的结果可以很好的筛选出一批真实的用户，对这些用户进行分析能得到较为可靠的研究结论。从而说明我们的方法在实际处理数据中具有可行性。

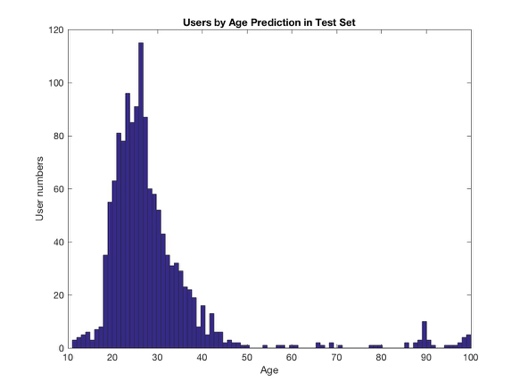
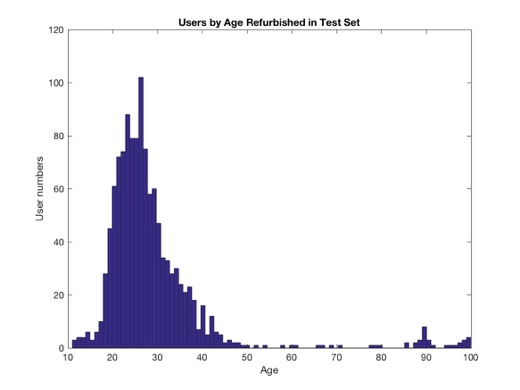
我们简单地分析了用户性别组成，用户年龄组成，用户平均收藏量，以及平均的评论数目。将未清洗过的数据，训练组内利用标记清洗过的数据，测试组内利用标记清洗过的数据，以及通过算法预测恶意用户再进行清洗过的数据进行对比，每一种指标一共四张图，自左上到右上到左下到右下依此排好。对比每一组内到四幅图，会发现第一张未清洗过数据的图和其他张图有比较明显的差异，而其他三张图之间差异则非常小。这说明了：第一、恶意用户与非恶意用户之间的使用行为存在比较明显的差异，考虑到恶意用户的数据不能反映真实使用行为，因此在进行数据分析的时候有必要剔除掉这部分用户的数据，才能得到较为可靠的分析结果。第二、第二张图与第三张图差异很小说明给定的数据是高度分散的，每一组我们的分组并没有破坏这种各向同性，从而每一个分组都能很好地代表样本全体，因而对第一组样本进行训练得到的模型也能很好地反映整体样本情况，可以很好地用在第二组样本上。第三、第三张图与第四张图差异很小验证了前面第一组样本具有很好的代表性，同时也说明了我们训练得到的模型具有很好的使用性，虽然简单粗糙，但已经能够在用于真实的用户数据挖掘中，进行清洗。



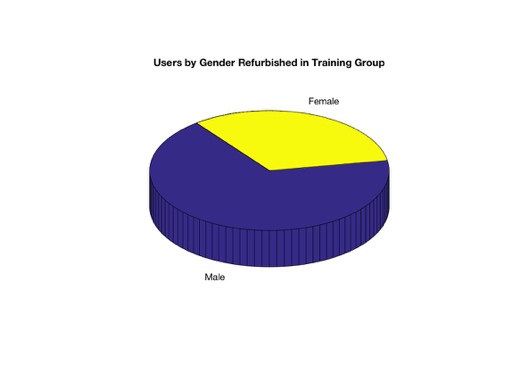
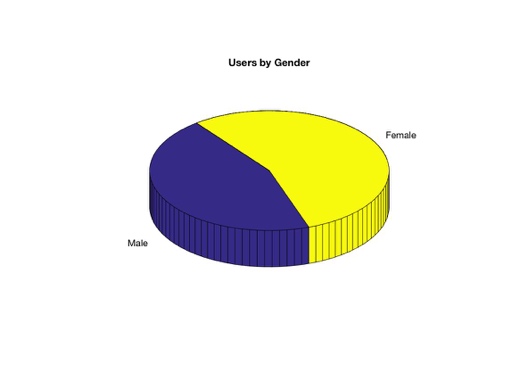


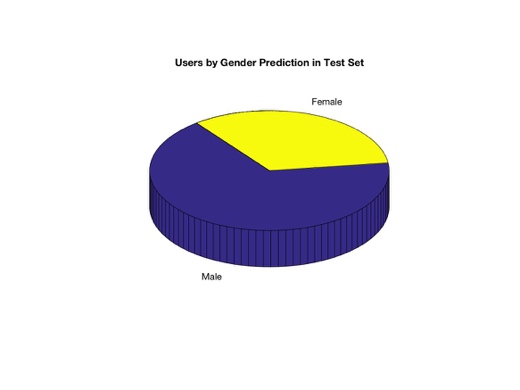
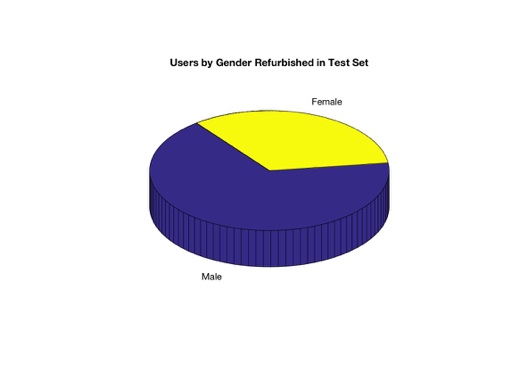
**（**图13不同数据组年龄分布统计对比**）**

****

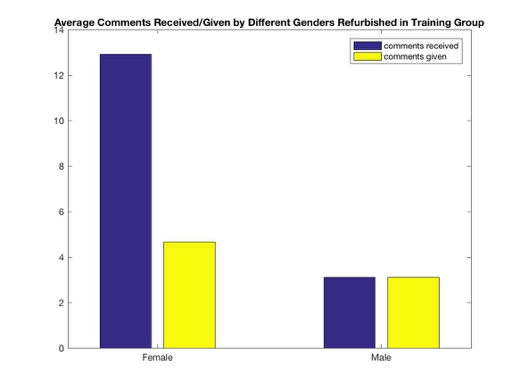
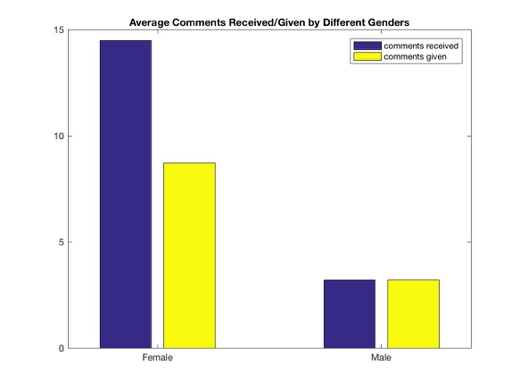
****

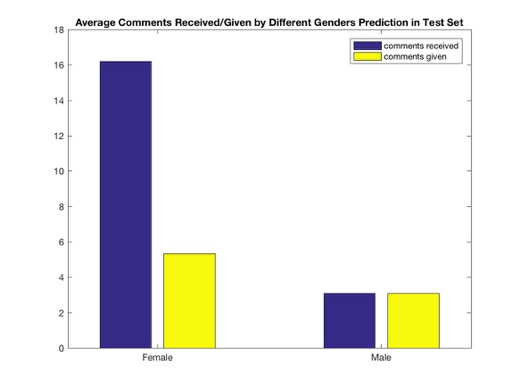
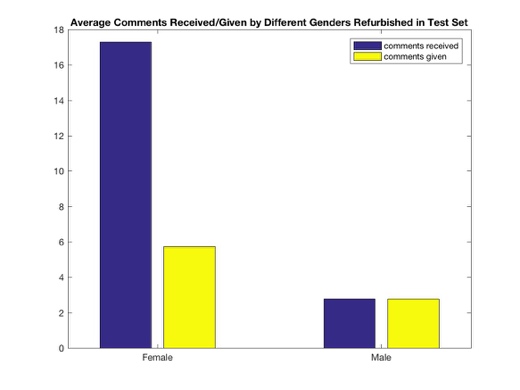
**（**图14 不同数据组内年龄分布直方图**）**



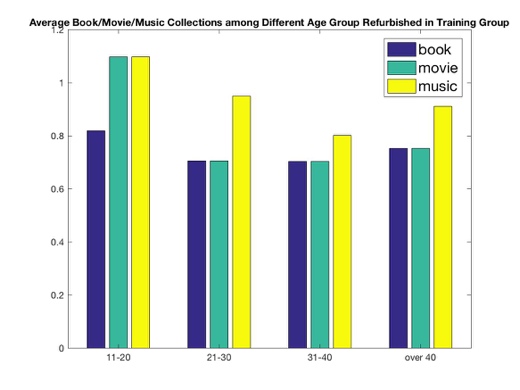
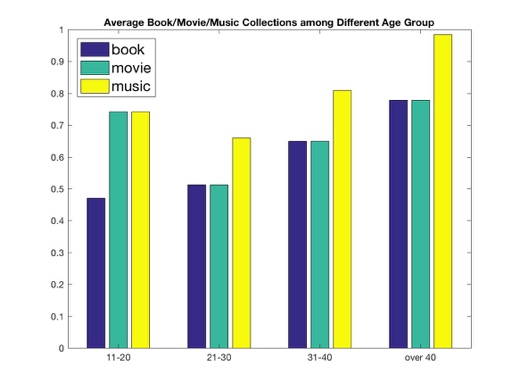


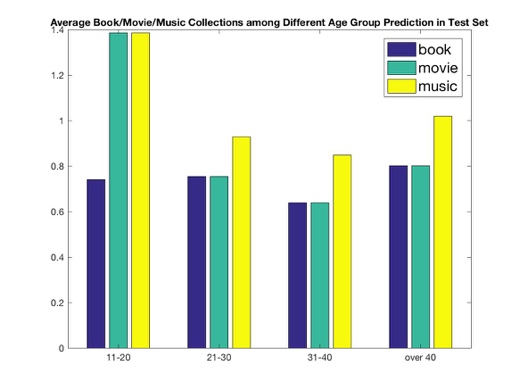
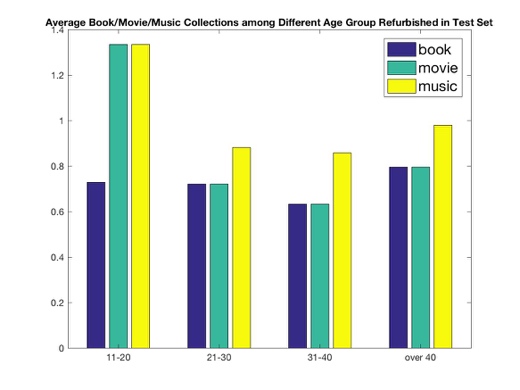
（图15 不同数据组内用户性别分布）





（图16 不同数据组内不同性别用户评论分布）





（图17 不同数据组收藏量随年龄分布统计）

1. 总结

在本次大作业中，我们先研究了恶意用户的一些身份和行为特征，定性地判断了哪些群体中恶意用户的比例较高，对个人用户甄别垃圾用户提供了一个判断的参考，有一定的意义。紧接着对于海量数据的研究，我们认为这种方法并不可行，准确率较低，具有很高的“误杀率”，为了解决区分恶意用户与非恶意用户这个问题，我们使用了机器学习的现成算法来完成分类工作，并使用了MATLAB2016版本自带的Classification Learner交互式界面完成了机器学习的工作，学习并训练产生了准确率为90%左右的模型，并且输出了此模型的代码，生成trainClassifier函数进行之后的分析。完成了这项工作后，我们用测试组验证了训练出来等模型等可靠性，并且使用此模型清洗了测试组数据，并做了一些简单的分析，由于时间和精力有限，我们并没有广泛地研究数据中有趣的一些特征，毕竟可以分析的东西实在是太多了，我们的重点工作放在了对恶意用户的分析和建立可靠的模型上。但是在完成此作业的时候，我们注意到了一些男女用户，不同年龄用户的一些非常有意思的数据特征，在接下来的空余时间里，我们将进一步完成这部分的内容，使得我们对于陌陌用户的数据分析更加完整。