微博水军识别中期报告

## 数据采集

为了达到研究目的，我们需要在互联网上采集用户的微博个人信息及相应的微博数据来建立一个庞大的实验数据集，并在对其按微博事件进行分类，对其中的部分用户是否为水军用户进行标注。

为了获得用户的信息，我们编写了一个面向新浪微博的爬虫程序来爬取用户在微博中公开的用户信息以及相应的微博。以可能出现水军的特定微博事件为基础，我们在采集数据的过程中设置了关键词和发布时间作为过滤条件，爬取在指定时间段内发布的微博内容中含有相应关键词的微博作为某个微博事件的微博数据。在获取事件微博数据之后，我们以参与到事件当中的用户为种子用户，对用户截止事件起始前两个月内发布的微博进行采集，作为用户的个人数据集。通过这种方法，我们选取了10个可能含有水军的微博事件，总共采集了约85万微博用户的信息及其发布的微博。采集的数据如下图所示：

|  |  |
| --- | --- |
| 类别 | 数量 |
| 用户总数 | 853,041 |
| 微博总数 | 142,304,427 |
| 用户粉丝总数 | 130,334,187 |
| 用户关注总数 | 115,675,345 |

下一步，我们需要对数据集中的用户是否为水军进行定义和标注，在我们的研究中我们主要关注的水军类型为通过大量转发评论进行灌水的用户，这种类型的水军用户在微博中分布最为广泛影响也最大。传统的水军识别研究中，通常使用的标注方法是首先定义弱规则对用户数据进行过滤，比如用户微博中水军微博的比例等，在此基础上再进行抽样结合人工鉴定的方法进行数据的标注。这种方法的缺陷就是在大规模的数据集中使用这种取样方法会包含一定的偏差，而且这种偏差以及人工鉴别的差异性都是很难避免的。因此我们使用一种更为简单更为准确地标注方法。

我们获取标注的水军用户的途径主要通过在互联网上购买。我们通过在淘宝网等购物网站上购买水军账户，或是雇佣水军来发布灌水微博来收集标注的用户。具体实施过程中，首先我们在互联网上购买了约5万个左右的用来发布水军微博的微博账号，通过爬虫爬取各用户相应的微博信息。在另一方面我们在每个事件中随机选取10%的用户进行人工鉴定来进行真实用户的标注，相比于水军用户的人工标注，真实用户的人工标注难度相对较低，主观因素的影响也相对较小。标注过程结束之后我们便获得了总用户量约为4万的标注数据集，我们在接下来的研究部分将以此数据集或是其子集来建立训练集和测试集进行相关的实验与分析工作。

在完成对数据的标注之后，需要建立相应的训练集和测试集，对现有的多种水军识别算法进行测试。在接下来的第四部分，我们将首先对现有的水军规避进行分析，并使用采集的数据进行相应的实验。在此基础上，在第五部分我们将介绍我们根据之前的分析结合特定微博事件的场景提出的新的特征集合。在第六部分我们对新的特征集合进行实验与评估，并对特征集合中各类别特征的识别效果进行了实验分析，对各类特征进行评估并分析其对检测效果的影响。最后，在第七部分我们分析了现有研究的不足并对今后的研究工作进行了展望，在第八部分对本文进行了概括性的总结。

## 2. 规避策略的分析

在这一部分中，我们将对水军用户用来规避基于机器学习的水军识别方法的策略进行分析。在对这些策略进行深入了解之后，我们将通过实际案例与使用现有数据进行实验相结合的方法来对这些策略进行评估。

### 2.1 规避策略

在目前阶段，水军账号的持有者为了保证账号的长久使用通常都会采取一定的规避策略来躲避水军检测的识别。但是受到规避成本的限制，通常不会使用过于复杂的规避策略。现有的微博领域的规避策略主要有以下几类：1.通过修改账号的基本信息进行伪装。2.通过模拟真实用户的行为进行伪装。此外，还有部分通过对水军微博的内容进行修改来进行伪装的策略，由于出现频率较少本研究中便不做考虑。

#### 2.1.1 伪装账户信息的规避策略

通过用户的基本信息比如用户的粉丝数、是否认证、注册时间等等信息来判别用户是否为水军用户是最为常用的水军判别依据之一，因为这些信息通常直接体现了一个微博账户的声望或是活跃度，而且由于这类特征十分容易提取所以经常被用在基于机器学习的水军识别算法中。目前常用的基本信息特征规避策略包括增加粉丝数或关注数和修改用户昵称。

#### 2.1.2 模仿真实用户行为的规避策略

另外一种常用的疑似水军披露方法就是对水军账号的行为规律特征。就像在引言部分的介绍一样，本文研究的微博水军账号通常是通过在微博中发布大量的灌水微博来获取利益，所需要费用的多少基本上取决于需要发布的灌水微博的数量，而灌水微博的数量越多取得的效果也越明显。从这个角度分析水军微博账号的行为通常都具有十分明显的特征，比如许多时候需要一个账号进行频繁的转发评论等等。因此许多的研究便针对这些特征提出了识别策略，包括计算发布微博内容的相似度、发布重复的微博的数量以及一些类似的特征。为了避免这一类的特征，水军通常采取的策略有避免同一账号多次转发和增加日常微博行为两种。

### 2.2 规避策略的评估

在此基础上，我们选择了几种现有的水军检测方法进行了实现，并在我们的数据集中进行了实验。通过分析实验结果中没有被识别出的水军，说明实际应用中确实已经有一部分水军成功的伪装成了真实账号避开了水军特征的检测。

## 3. 特征集合的选择

在这一部分我们将首先对特定微博事件进行定义，并阐述针对特定微博事件进行水军用户检测的需求和意义。在此基础上，我们提出了一系列水军检测特征，其中包括我们针对特定微博事件新提出的特征，也包含已有研究中被广泛使用的检测特征。具体的特征包含属性特征、行为特征、内容特征三类。在接下来的部分我们将对这三类特征进行详细的描述与解释。

### 3.1 行为特征

行为特征，从一定意义上来说描述的是一个用户对微博的使用习惯，比如用户习惯在每天的特定时间登录、习惯每隔一个相对稳定的时间发布一条微博、习惯使用某种手机客户端手机或是某种浏览器等等。而对于水军来说，发布大量的灌水微博很明显属于一种十分反常的微博行为，不符合普遍的微博用户使用习惯，所以我们可以针对水军的一些行为特点提取特征进行检测。此外为了对特征的选取辅助说明，我们随机选取了某微博事件中标注的1000名真实用户与随机选取的1200名水军用户组成了辅助数据集，来完成对特征显著性的部分说明。

#### 3.1.1 发帖频率

真实用户随着微博使用时间的增长，发布微博的频率会趋近于稳定在一个符合用户微博习惯的频率，尤其是现阶段越来越多的水军倾向增加转发账号减少单位账号转发数量，真实用户与水军用户的发帖频率的差距也越来越明显。

在实际计算过程中为了保证数据的时效性，我们统计用户最近2个月内发布微博的平均每小时的发帖数，所以用户$v$的发帖平率$F\_{post}$可以表示为：

$$F\_{post}= \frac{N\_{post-dura}(v)}{\left|T\_{dura}\right|}$$

其中$T\_{dura}$表示计算发帖频率的时间段（以小时为单位），$N\_{post-dura}$表示在指定时间段内发布的微博数。为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：



如图所示真实用户与水军用户的发帖频率有着明显的差异，不仅是真实用户的发帖频率要相对高于水军用户。而且从曲线中看出水军用户的曲线呈有规律的阶梯状，这表示水军用户的发帖频率集中在某个较小的区间；而真实用户则是平滑的曲线，进一步说明了两类用户在发帖频率上的区别。

#### 3.1.2 上网方式数与转发微博上网方式数

随着新浪微博等微博在生活中的日益普及，人们在浏览微博的方式上也有了更多的选择，无论是PC端的各式浏览器、微博客户端还是移动端的微博app， 在这里我们定义用户访问并发布微博的途径为用户上网方式。对于普通的用户来说，更换访问微博的方式是十分常见的，比如在外出行会使用手机客户端而如果在电脑跟前则会直接通过网页访问。相比之下，水军账号由于通常都是通过转发器或是微博api大规模发布微博，所以水军用户与真实用户使用的上网方式种类存在明显差异。

为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：



从图中可以明显看出水军用户与真实用户在常用上网方式上的区别，整体上真实用户使用的上网方式要丰富很多，而水军用户使用的相对单调。

#### 3.1.3 是否参与热门微博

 对于通过频繁转发或评论来进行灌水地微博水军来说，需要为他们指定一条或几条目标微博，而他们需要做的就是对这些指定的微博进行转发及评论。综合我们对水军服务提供者的调研，我们定义转发数或评论数大于100的微博为热门微博。相应的，以100次转发或评论为分界点，对非热门微博进行转发或评论的用户是水军用户的可能性通常很小。

### 3.2 内容特征

 内容特征是最为常用的水军识别特征，因为无论水军如何伪装，最终都要通过发布的微博内容来传达相应的信息，所以水军发布的内容进行研究是识别水军的必备特征。但是考虑到微博中水军有出现的情况日益复杂，比如某一条微博被水军大量转发，有可能是为了对其中的产品进行推销，也有可能是为了扩大博主的影响力，所以我们放弃了对微博内容的语意研究。我们选择对用户的微博内容进行整体的研究，选取了以下三个特征对内容进行描述，避免了对内容进行语义分析的高计算复杂度。

#### 3.2.1 原创微博比例与转发微博比例

对于水军账号发布的微博，其中原创微博和转发微博的比例具有明显的特点。从水军微博的行为分析我们可以发现，大部分的水军都是以转发或评论他人的微博为主，而原创微博的比例则相对很低。

根据之前的描述，我们设用户发布的总微博数为$N\_{all}$，其中原创微博数为$N\_{ori}$，转发微博数为$N\_{rp}$，在此基础上可以计算一个微博用户的原创微博比例$R\_{ori}$和转发微博比例$R\_{rp}$：

$$R\_{ori}= \frac{N\_{ori}}{N\_{all}}$$

$$R\_{rp}= \frac{N\_{rp}}{N\_{all}}$$

为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：



#### 3.2.2 转发微博平均“@”符号数

我们认为可以使用“@”符号来描述用户的多次转发特征。根据我们的对水军服务提供者的调研，水军进行灌水地一般过程是用户指定或自己发布目标微博，水军账号再对目标微博进行转发。为了节约时间和操作成本，不存在通过不同的用户对原文进行间接转发或二次转发的情况。也就是说，水军用户在转发时包含附加信息中的“@”符号数通常为0，所以我们针对这个特点提出了用户转发微博中平均出现的“@”符号这一特征。假定用户$v$的转发平均“@”符号数可以表示为$A\_{rp-at}$，则可以通过如下公式进行计算：

$$A\_{rp-at}\left(v\right)= \frac{1}{\left|N\_{rp}(v)\right|} ∙ \sum\_{u\in N\_{rp}(v)}^{}N\_{rp-at}(u)$$

其中$N\_{rp}$表示用户的转发微博总数，$N\_{rp-at}$表示用户转发微博附加内容中的“@”符号数。为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：



### 3.3 属性特征

 尽管越来越多的水军持有者已经意识到了对水军账号的基本信息进行伪装的重要性，但是我们还是将属性特征作为一类重要的特征加入到了我们的特征集合当中，这样的做法主要是出于以下几方面的考虑：1.并非所有的水军持有者都会去进行帐号个人信息的伪装，尤其是小型的水军团体通常还是会使用比较原始的灌水方式和未经过伪装的账号，这样可以一定程度上提高系统的召回率；2.考虑在实际的应用过程中属性特征的计算复杂度更低更加便于提取和使用。

#### 3.3.1 粉丝数与关注数

在之前的部分我们也已经介绍过了用户的粉丝数（关注数）这一概念，这一特征可以体现用户在微博中的声望或受关注程度，也是用户活跃度的一个侧面体现。粉丝数（关注数）这一特征可以直接使用微博中给出的数字，计算复杂度低。为了确定这一特征的加入是否具有意义，我们在我们标注的数据中随机抽取真实用户与水军用户共2000人，对这两类用的粉丝数（关注数）分布进行了研究如下图所示：



从图中我们可以看出，尽管水军持有者可以增加他们持有的水军账号的粉丝数（关注数），但是水军账号的粉丝数（关注数）的整体分布仍然与真实用户有所区别，这说明增加粉丝数（关注数）的策略仅仅规避了基于数值的检测条件，并不能完全规避属性特征的检测。

#### 3.3.2 粉丝关注比

为了计算这一特征，首先我们要提取用户的粉丝数和关注数，这里我们这两个特征表示为$N\_{fans}$和$N\_{follows}$。用户的粉丝关注比代表的就是用户的粉丝数与关注数的比值，所以用户的粉丝关注比$R\_{fafo}$可以通过以下公式计算获得：

$$R\_{fafo}= \frac{N\_{fans}}{N\_{follows}}$$

选择这个特征的原因是因为在我们调研的过程中发现，大部分提供水军服务的团体或个人通常都不仅限于提供微博水军，而是包含多种有关新浪微博的业务，其中一个重要的部分就是为用户增加粉丝，也正是由于这个原因大部分的水军账号的粉丝数是要远远小于关注数的。相比之下普通用户的粉丝数和关注数比较接近，甚至对于一些大v用户粉丝数要远远大于关注数。为了对选取特征的显著性进行检验，我们使用我们采集的数据对水军用户与真实用户的这一特征分布进行了分析，如下图所示：

