– Sina Weibo –

微博反作弊原理详解

EnXu
Sina

2015 年 6 月 25 日

1 业务背景

[1] 微博是一个快速的公众信息传播平台，当我们去浏览微博的时候更希望看到的是对自己而言有用有效的信息，但是任何事物都有其两面性，微博上的信息同样良莠不齐，如果不对微博中的垃圾数据进行清洗，可以想象浏览一个垃圾信息充斥的微博就像走在一个遍地狗皮小广告的大街上一样令人难受，也因为此需要对垃圾微博进行处理。

首先来明确下垃圾微博的分类，由于没有明确的标准说明什么是垃圾微博，这里简要的说下垃圾微博有哪些类别，针对不同的目的垃圾微博可以分为**色情及敏感类微博、无关信息微博、低质信息微博** 等几大类，其中作弊需要处理的就是无关信息类的微博，本类别的具体表现就是**文不对题**和**图文无关**两种现象(具体表现见章节2)。针对这两种微博的现象的处理就是我们所说的微博反作弊(weibo-antispam)[2]。

目前微博的反作弊还处在探索阶段，业界也没有成型的理论以及工作可以做，而且针对具体的业务也有着千差万别，就目前微博的反作弊而言，要想广泛的全面的定位内容无关微博是困难的，目前做到的是针对线上作弊量较大的广告，段子以及导流类别的微博特定性的进行了处理。另外这里说下微博反作弊目前面临的困难：

时效性问题 判断微博是否作弊要求及时，否则过时判别就会变得无意义。由于计算机没有世界观，所以判断一个内容的相关性经常依赖于预先进行的相关统计，而当

[1] 感谢身边的同事的帮助，没有你们的工作这些也无法完成，这里特别感谢庆涛，谢谢他一年来在反作弊上默默的付出这里的很多数据都与他的工作离不开

[2] 至于为什么称作是微博反作弊原因在于微博的反作弊重在内容不像普通作弊重在用户行为

一个作弊者用一个突发事件来作弊，这时我们是很难立即找到相关的事件进行判别的，因此很难对一个微博立即进行作弊判别。

判别的困难性 判断作弊需要判断一个短语与另一个短文本之间的相关性，而这本身就是困难的，例如名星以及表情这些泛词，一个名星涉及生活的方方面面，与他相关的内容广泛，即使我们能够存储一个明星的所有的关联关系内容，却无法找到合适的办法判别内容是否就是以这个为主题，既他是不是文章的主题。另外还有图文不相关则更加难于判别，这涉及图片语义问题。

短文本处理的困难 业界目前公认的短文本最难于处理再加上微博内容本身的非规范性与随意更加难于处理。如微博中充斥着大量的火星文，有话不好好说等现象，所谓一千个人就有一千个写法在微博上表现最为明显!

作弊手段的多样性 作弊是与人斗，所谓道高一尺魔高一丈，当我们找到一种作弊手段加以打击后就会出现另外一种更加嚣张的作弊手段，让人防不胜防。微博是某些人的生命线，作弊更是他们生存的重要手段，他们经常会恶意变换内容干扰作弊判别等，如一个微信号的写法就高达30种以上，这些人也是蛮拼的!

作弊需与时俱进 随着微博的发展作弊的手段会发生转移这就需要作弊也随着转移，如现在一些段子作弊转向视频内容作弊，这点很像警察与罪犯的故事，是个永远不完的猫捉老鼠游戏。

反作弊上虽然存在着种种困难，但是也并非全不可行，我们可以由简到繁先多后少的一步步进行，经过我们反作弊的一段时间工作，目前线上的主要作弊手法也已变成较难识别。最后对于从事这项工作有几点建议就是：**步步蚕食、勿以善小而不为、随机应变 、预测是最好的防御、作弊的本质是人在作弊**。反作弊路上漫漫其修远兮，吾将上下而求索，以此与诸君共勉之。

2 作弊初窥

兵法云知己知彼百战不殆，了解作弊者的动机以及方法手段有助于我们针对性的下药，以便做到见药起效。一般而言广泛性的作弊指的是所有的文不对题以及图文无关的微博，但要做到绝对的识别是困难的且不必要的，而且很多的微博本身就是文不对题而且也很模糊，退而求其次这里更加关注那些带着目的去作弊的微博，目前微博的作弊者常见的作弊目的无非以下几种，包括有*养号*、*求关注*、*导流*、*推广*以及部分*无意识作弊*。作弊者惯用的手段就是利用一个亮点(HotSpot)来吸引用户同时附上自己收益点(Profit)以期达到利益最大化，目前微博里吸引用户的亮点包括*热门话题*、*热门名星*、*热搜榜* 这三种，举个例子一个卖衣服的作弊者经常就会在热门话题发表一些自己所卖东西的简介以期获得较大用户流量，如图2a。值得注意的是这里的作弊者不但指的是人，有可能也是某些机器，如一些恶意APP也会作弊，其中导流类的作

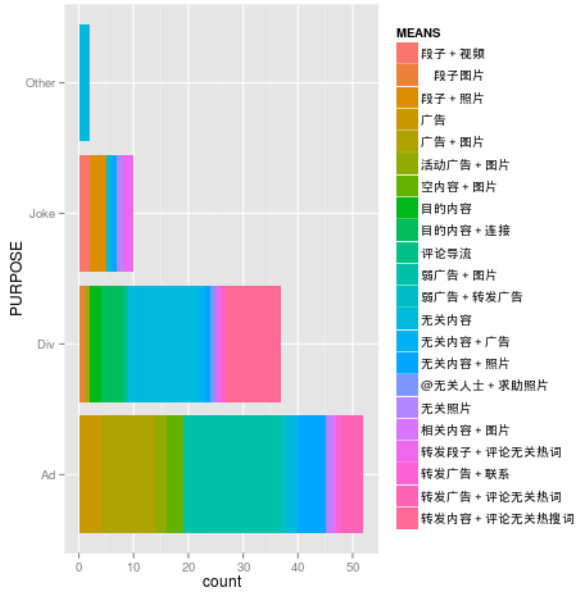


Figure 1: 作弊分布

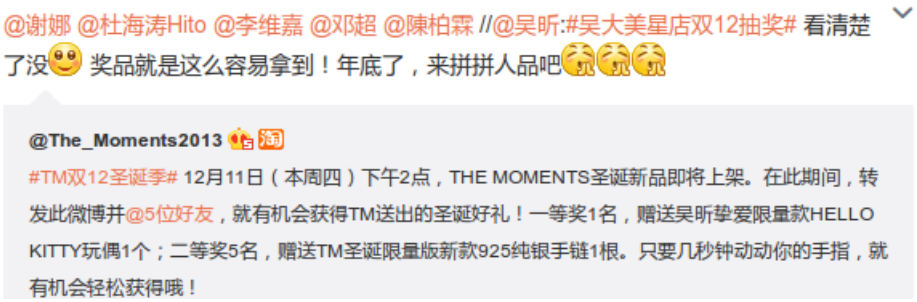
弊APP就占据了很大一部分。具体的一些作弊例子参见图2，图1是其中某一次线上作弊手段调查的分布情况，到目前为止微博中作弊最严重的类型依然属于广告。

以上所述内容均是针对某一次具体作弊微博而言，下面说下作弊用户(**恶意利用微博作弊达到某种目的的用户**)，如果就一个作弊用户而言，其行为是一系列的微博集合体。作弊用户所表现的形式要更加的复杂，有些作弊用户时好时坏，有些彻头彻尾些，有些则隐藏的很深，凡此种种不一而至。全面的客观评价一个用户是否作弊需要结合其长期的历史行为数据来判断，如一些作弊用户经常的发很多热门话题，而且这些话题所讨论的内容经常与这些话题本身毫无联系。作弊用户的作弊手段经常会发生变化的，他们会采用上述一切可能所有的手段来进行作弊，同时某些用户会伪装的发一些正常的微博以干扰判断。举个典型的例子用户，一般而言作弊用户是穿插性的发送些作弊微博，在整个行为历史中作弊者的作弊微博一般较正常人多，他们往往都是利用这些作弊微博吸引用户进入他们的微博查看与他相关的微博，但也有些少数作弊者本身微博不作弊但利用买粉帮助他们进行导流，使得用户进入其相关微博。

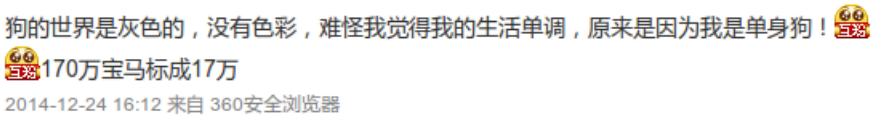
(a) 话题内容不相关

(b) 图文无关

(c) 多个无关话题



(d) 转发作弊



(e) 穿插内容作弊

Figure 2: 作弊示例

3 打击作弊

前面我们将作弊分成了两种情况对待，一种是针对单个微博一种是针对用户，按照作用对象**反作弊分为页面反作弊以及用户反作弊**，形象的说两者的关系是治标与治本的关系，页面级别治标头痛医头脚痛医脚，哪个微博作弊了就打击那个好处是快速精确但是这种打击方式没有把握作弊是人在作弊的本质因此会疲于应付，用户级别治本除去根源，直接屏蔽指定的用户好处是能够更加长远的打击作弊行为但是会误伤很多微博见效慢且这种打击只能针对长期的有目的性的作弊不能够针对突发作弊状况，所以只有两者结合标本兼治才会有良好的效果,以下分就介绍两种作弊处理方式。

3.1 页面反作弊

页面反作弊是直接针对一条微博作弊现象而处理的，那么如何判断一条微博是否作弊，严格的依照定义就是图文无关以及内容无关两种，然而现实总是残酷的，依照目前的水平我们无法做到图文无关以及内容无关判别，即使是人在某些领域由于信息不完全也做不到如此要求，也因此需要另外寻求解决之道，在这里借助概率知识判别，判断一条微博是作弊的概率 $P_c(Doc)$,考虑到微博是由某个用户推送的，因而一条微博作弊的概率是 $P_c(Doc, Person)$,如果考虑用户行为的时间性，那么作弊概率变为 $P_c(Doc, Person|Time)$, 整个反作弊系统的运行原理就是求一条微博的 P_c 。

目前反作弊假定的是用户行为不变性，既一个用户的行为不随着时间的变化而发生变化，在此基础上求解 $P_c(Doc, Person)$的。由于 $P_c(Doc, Person) = P_c(Person) \cdot P_c(Doc|Person)$，由此也可以看出反作弊体系需要两种判别体系，一个是判别用户作弊的 $P_c(Person)$即用户级作弊判别，一个是利用用户信息判别微博作弊的体系 $P_c(Doc|Person)$，这也就是页面反作弊的目的。

以上从原理上说明了整个反作弊体系与页面反作弊的目的，下面说下对于页面反作弊的具体的一些求解。目前反作弊页面模型利用的是分类体系，具体过程是 $P_c(Doc|Person) = P_c(f_1, f_2, \ldots, f_n|Person = \mathcal{C}, Model = \mathfrak{M})$, 换句话说页面级别反作弊在给定的模型情况下，利用用户人群的分类，结合微博现有的特征来判别是否作弊的，这里的f_i表示微博的特征，如热搜词数，昵称数等,$Person = \mathcal{C}$是用户的分类信息，由用户反作弊提供或者其他来源提供，例如用户是一个广告或者粉丝用户等等，这里的模型目前选取的是决策树类的策略，当然以后可以发展为自动学习的其他机器学习模型。 以上都是些理论叙述，目的在于阐述页面反作弊的数学原理供反作弊上的工作指导以及日后的改进方向提供参考。下面说下工程上如何来实现上述的计算过程，页面反作弊的整体框架参见 3,这里先简单介绍下框架的几大部件，页面反作弊的主要模块包含：容器,控制器，基础服务，输入源，特征模块，策略模块，输出源七大模块

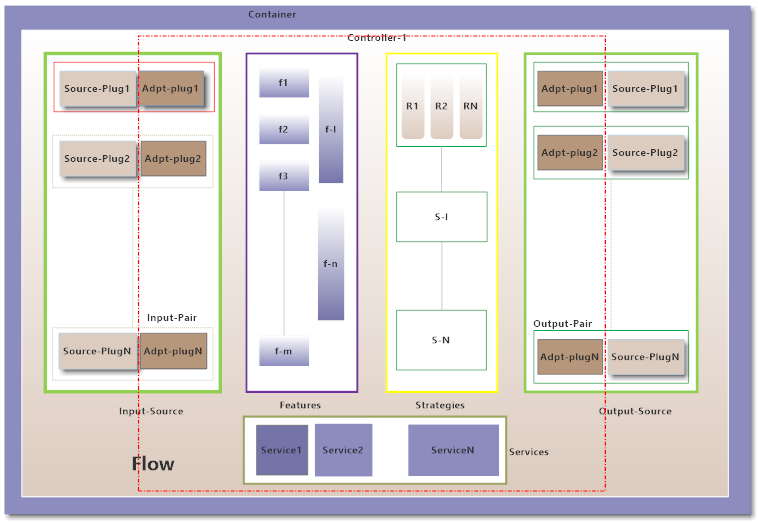


Figure 3: 页面反作弊框架

容器 提供反作弊的运行环境，控制反作弊整体的运行参数，目前只是采用简单的控制台环境，以后可以发展采用web容器甚至zookeeper(分布式协同)，方便参数修改以及控制监控反作弊的运行状况

控制器 控制反作弊剩余模块的运行，协调反作弊各个模块的处理流程，维护各个模块的状况，目前反作弊提供了3种控制器，其中最复杂的控制器大量采用队列缓冲与多进程处理提高执行效率以及支持降级处理，不同的控制适用的场景不同，反作弊针对离线测试、在线测试、线上生产、被动调用都有相应的控制器

输入源 输入源由一系列的（数据读取插件，解码插件）对组成，该模块功能的是读取指定的源数据并将其转换为标准输入格式，另外解码器另外一个功能是过滤掉不适合反作弊处理的数据（失效，非法），现有的输入插件支持队列，文件夹，文本以及数据库，解码插件支持标准@格式，json格式

基础服务模块 由一系列基础服务组成的插件，服务于特征插件以及策略插件，提供计算所需的基础功能，这些服务可以是本地字典亦或是外部调用等

特征模块 由一系列特征插件组成，各个插件的功能是计算对应微博对应特征的数值信息(之所以选择数值是为了便于后续的计算)与描述信息(便于信息共享减少重复计算)，特征插件之间可以单向依赖，由控制器根据需要自动感知调用依赖的特征计算。

策略模块 根据微博的特征向量数据，匹配对应的策略，不同类型的丛向策略(如低质策略与反作弊策略的区别)都会被计算，而同一个类型的横向策略(反作弊的策略的各个子策略)一旦其中一条被匹配就会被返回。这里的策略可以是人工指定的规则亦或是各种规则树，也可以是各种自动分类的数学模型。

输出源 功能与输入源相反，是将反作弊产生的数据转换为需求格式并输出到指定的源头，反作弊可以同时支持N条输出流，每个输出流都可以有自己业务的需求，反作弊目前支持线上作弊标记输出外，还支持着作弊日志输出(用户行为特征日志),相关词优质内容微博输出(用于分词器改进以及相关词等机器学习在线语料)

反作弊框架里特征模块以及策略模块既是求解上述给定的 $P_c(Doc|Person)$，首先由特征模块提取微博的特征再由策略模块结合具体的模型策略判别是否作弊。在目前的反作弊框架下，根据上述的模型求解过程可以知道反作弊的准确程度取决于特征的准确以及细化程度，用户分类的准确度以及模型（策略）的适应范围，也因此反作弊的大量工作就是在做这些的改进工作，整体上反作弊的工作都是在分析数据，在分析这些数据的基础上决定是提高特征准确率亦或是增加特征亦或是改进策略模型，至于如何分析数据，如何提高模型准确率，如何改善特征在这里不再详述。（这方面请参考weka,R,数据分析，机器学习等其他书籍）

另外顺便提下开始时说过反作弊的工作最好能够主动防御，对于页面级别而言主动防御有两个地方可以做到，一是提高模型的泛化能力，二是特征自学习，三是细化用户类型针对不同的用户采取不同的对策。顺便说下对于反作弊框架而言其发展应该是广泛的数据质量控制，不限于反作弊，低质垃圾等数据质量鉴定也可以在此完成。

3.2 用户反作弊

用户反作弊针对的是恶意作弊的用户，对于临时性的突发作弊用户由页面级别处理，在页面反作弊的时候提到用户反作弊的目的是判别用户的作弊概率 $P_c(Person)$,目前的反作弊框架下用户作弊概率 $P_c(Person) = P_c(T_1, T_2, \ldots, T_n)$,其中$T_i$是该用户行为的统计量 $T_i(C_{t_1}, C_{t_2}, \ldots, C_{t_n})$, C_{t_i}表示用户在 t_i时刻的用户行为，我们将一条微博看作是发这条微博的用户的一次行为，因此是一次用户行为就是一条微博的所有特征数据组合 $C_{t_i} = (f_1, \ldots, f_n)_{t_i}^{doc}$。某个用户在一段时间内的微博就汇集成了该用户在一段时间内的行为数据，这些特征数据可以由页面反作弊提供。

上述内容可以看出用户反作弊的主要工作就是寻找合适的统计量，调整作弊识别模型，至于特征数据的完善由页面级别完成，与页面级别一样工作内容仍然围绕分析数据展开。对于用户级别的架构其整体上与页面级别相同，只是页面级别的特征在这里是特征的统计量，用户级别的识别好坏除却特征本身的建设外另一个重要的方面就

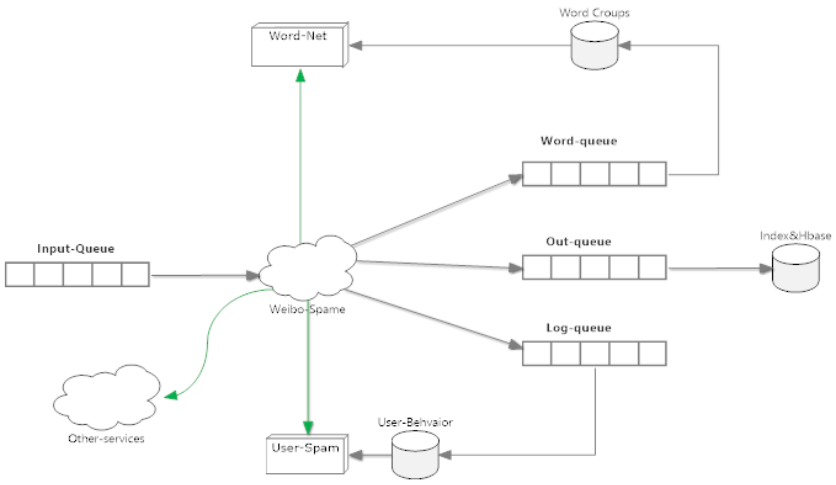


Figure 4: 反作弊数据流程

是如何找到好的统计量。目前反作弊的统计量大多都是均值统计量，对于方差，熵等其他的统计量使用较少，这些都是日后需要改进的地方。

以上所说的用户级别反作弊是基于**用户行为的**,另外一种反作弊是基于**社交关系**即根据用户之间的关联关系判别用户属性的方法，微博里面可以使用的关联关系有**转、评、赞、关注、被关注、@用户、链接、共同话题，共同联系方式**，这些都是反作弊可以是用的社交行为。而基于社交关系的用户行为作弊属于用户反作弊的另外一个辅助改进方向，其主要是利用图挖掘技术相关内容请参考具体书籍。对于用户级别反作弊的发展与页面级别一样不限于反作弊，在用户兴趣挖掘，人群定位上也可以发挥作用。

3.3 系统协调

要让用户级别与页面级别的处理工作起到相辅相成的作用，需要找到两者的结合点。依据两者的计算目的，不难看出页面级别的工作为用户级汇集用户行为数据，用户级别的工作为页面级别提供用户分类信息。实际上系统的整体工作也是这样，具体处理流程参见图 4 。在实际系统里面反作弊有着大量的周边工作需要完成，这些都是反作弊系统的辅助改进工具，如作弊联系方式挖掘，广告词挖掘，作弊数据分析工具，case分析工具，人工干预工具等等，所有的这一切都是需要完善的，当辅助开发工具，辅助分析工具，辅助调整工具等都完善后系统才能算得上是一个完善的成熟的体系。

4 关于工作

个人认为反作弊是微博的第二道防火墙，一个数据层面的防火墙，这项工作与微博的关系就像安保与住宅，平时大家不会在意可是问题来时就会很严重，因此工作会使人陷入做好了平常无事不关你的事无存在感，来事就是你的错无辜感，容易陷入痛苦的循环，引用主任的一句话，"我们是弱势群体!" [3]。反作弊的工作本身依赖于很多的基础数据工作，初期我们更加应该完善的是基础体系的建设，如分词体系，类别词词库体系，相关词体系，行为库等的基础建设，没有这些基础工作很难开展后续的工作，工作本身涉及的技术体系包含，自然语言处理，图像处理，数据分析和机器学习等几个部分，这几个技术都是反作弊中广泛使用的技术，日常工作需要注重对这些技术的积累以及人员积累。

当反作弊进入中期时反作弊的主要就是完善微博基础特征以及用户行为特征，同时实现两者之间的互通互惠的封闭体系，微博特征识别作弊特征的同时服务于用户行为特征，反过来用户行为数据除却进行作弊用户识别外还可以辅助页面作弊进行识别。这期间除却这些基础工作外更主要的是各项识别策略的迭代，这是一项长期以及繁杂的任务。本阶段的主要任务是数据的收集整理分析，模型改进调整，这些工作需要的是数据分析人员的耐心和发散思维。

完成了反作弊中期后反作弊体系基本已经成型和稳定了，但是这还只是个单纯的城堡式防御体系还没有很大的防御能力，前面说道最好的防御是预测，剩下的任务就是如何预测某个用户的行为，该阶段也需要分三阶段走。第一步是静态预测，此阶段只是简单的根据历史统计进行作弊判断，并不涉及用户的状态转换等行为考虑，该阶段主要是识别那些明显的行为固定的作弊用户，第二步是动态预测，这步将进行用户行为状态转换记录，此时用户的行为不在是一系列静态的点而是一个马尔科夫过程，在这一阶段识别那些飘忽不定难于防御的作弊用户，第三步是基于用户互动以及社交关系的高级预测这步可以在第一步中做一些，但是彻底的完善还需要系统具备完善的用户关系记录，需要基础平台的完善。

当然展望更远的未来，反作弊长足的改进还需要人工智能技术的提高，有一天反作弊也可以达到像《疑犯追踪》里的Samaritan那样的高度。

[3]满纸荒唐言，一把辛酸泪，望各位M们勿介意!