从Pins, Tweets 研究用户通过图片和文字在社交网络里的行为

RaphaelOttoni† rapha@dcc.ufmg.br

DiegoLasCasas† diegolascasas@ufmg.br

Jo˜aoPauloPesce† jpesce@dcc.ufmg.br

WagnerMeiraJr.† meira@dcc.ufmg.br

ChristoWilson‡ cbw@ccs.neu.edu

AlanMislove‡ amislove@ccs.neu.edu

VirgilioAlmeida† virgilio@dcc.ufmg.br

†Universidade Federal de Minas Gerais, Brazil ‡Northeastern University, USA

**摘要**

如今，人们通常在多个社交网站(OSN)上有自己的账号。类如，最近的一项研究表明

34%的Twitter用户也同时也有Pinterest的账号。这种情况引出了一个有趣的问题，比如说：用户是否同时在两个社交网站发表了同样的内容。另外，在这种情况下，信息的来源是哪里？鉴于社交网络之间的相互关系，如果不彻底分析多个网站的活动就可能混淆信息传播的真实状况。

在这项研究中，我们的第一步是对一个跨多个社交网站的用户的完整学习。我们收集了30000个同时拥有Twitter和Pinterest帐户的用户，爬取了他们近三个月来的所有信息和每天的活动。我们开发了一种新的方法比较用户在两个网站之间的活动。我们发现了一个通过比较两个网站发现了一个全局的模式，用户总是趋向于先发布在Pinterest上，然后再发布在Twitter上。我们的发现能用于引导每一个网站上对用户行为的学习，以及对整个社会网络共享动态的研究的理解。

**1 Introduction**

在线社交网络(OSNs)现在是一种流行的联系、交流、分享的方式，现在好多的互联网门户网站都在为上百万的用户服务。因为这些网站的大量普及，大量的用户数据和用户交流的内容为解释人们社交的模式提供了前所无有的机遇。

最近几年来，社交网站的数量和模式都已经超越了传统意义上的社交网络，比如说Facebook。其中比较流行的还有 Tumblr (短的博客),Pinterest(分享收集的照片), 还有FourSquare(分享他们的位置信息)。实际上，许多用户都同时活跃在多个社交网上(最近的数据表明超过34%的twitter用户也同时拥有Pinterest的账户(Pew Research Center 2013))。虽然现在已经有许多解释用户在社交网络上行为的研究。但是还不清楚用户如何在不同网站间分配自己的时间。比如说，用户是否在不同的网站上分享了相同的内容，这些内容的来源是什么等等。

在本文中，我们就来仔细研究在不同网络见用户的行为，重点对热门网站Twitter和Pinterest进行研究。Pinterest是一个照片共享的网站，它允许用户共享他人的照片合集（”boards”）Pinterest的主要设计是能为用户提供分享来自其他网站的照片(“pin”)的功能(它不像网站如Flickr,这是围绕着用户自己上传的图片)。Twitter和Pinterest的用户能够“关注”别的用户，这意味这其他用户分享的照片会在用户登录时转发给用户。

我们的目标是理解用户在这两个不同网络间的活动。首先，我们第一个目标是了解用户是否在这两个网络间共享了同一个身份。其次，我们调查用户在这两个网络间不同的兴趣点。在Pinterest上，用户的兴趣能够通过包括在不同的圈子里分享和转发照片推测，其中的每一张照片都被标记为一个类别。在Twitter上，用户的兴趣点能够通过他们的每一条推(tweets)来推测。问题是我们的目标是回答，用户首先关注的是什么，Twitter还是Pinterest?一个用户同时在两个网站上分享了同样的内容吗？怎么才能通过比较图像和微博来发现他们的共同兴趣。

在比较Twitter和Pinterest时，我们面临着一系列的挑战。不同于Twitter的是，Pinterest是具有挑战性的研究，因为上面最为共享的内容是图片（而不是文本），并且没有为公众提供的API(应用程序通用接口)来收集数据。我们开发了一个用于收集Pinterest数据的程序，并且收集了大约30000同时拥有Twitter和Pinterest账户的用户数据。我们下载了他们所有的tweets和pins，结果是大约290万的pins和710万的tweets.此外，我们开发了一种新方法，它是能够比较基于图像的内容（Pinterest的）基于文本的内容（Twitter的）。我们所采用的内容类别，是由Pinterest的所提供的;我们证明了我们的方法来用来分类Twitter数据所达到的精度高于其他不同的方法。

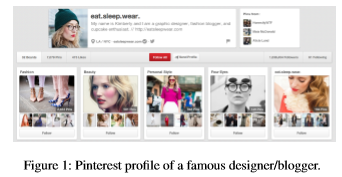
总体而言，我们的研究做出了三个贡献，第一，我们的目标是连接这些网站行为的宏观格局。我们发现，尽管考虑到了这两个网站上相同的用户群，我们还是看到了不同的行为模式。第二，我们用我们的分类方法来研究用户如何分配他们在不同网站的时间，我们发现用户在Pinterest上关注了更多类别的内容，相比于Twitter。但是Twitter的列别有更大的预测能力。最后，我们发现了一个新内容的传播往往是从Pinterest到Twitter, 这一结果表明，虽然Twitter是一个非常受欢迎的全球沟通的平台，但是像Pinterest这样的网站往往对新内容的产生起到至关重要的作用。

**2 背景**

**2.1 Pinterest**

Pinterest的是一个旨在让用户收集和共享分类的图像和录像的影像分享的社交网络。 Pinterest成立于2010年，并在2013年7月拥有7000万的用户群。它是目前世界上第27流行的网站，也是最快的吸引访问者通过搜索引擎访问的网站。（沃克2012）Pinterest的用户有80％的女性（Chafkin2012; Ottoni等2013），平均每月用户的使用时间是98分钟，这使得Pinterest成为第二使用率的社交网站，仅次于Facebook。 Pinterest成为了电子商务的重要渠道，最近的一项研究显示，Pinterest的用户价值近两倍于Facebook 用户。

在Pinterest上，用户创建一个个人的多板块的设置，没一个板块都有一个名字和描述（包含自由形式的文字和文件）就像类别一样。Pinterest预先定义了33个类别，从“女性时尚”和“美发”到“极客”和 “刺青“。这些基础的数据在Pinterest上是图像和视频。用户把他们发表到自己的板块里。每一条pins的特点包括一个相对时间戳（例如，“发布20小时前”，“发布于5天前”），一个描述（自由描述的文本），还有一个源的连接(如果它来自于一个第三方网站)。Pinterest唯一支持的直接操作是repin(转发)，转发的理由，还有是否喜欢。默认情况下，Pinterest的所有数据是公开的。



Pinterest在社交上的结构和功能很类似于Twitter.个人的设置中，包括一个图片设置，一个简短的自我介绍，还有他的板块列表，发过的pins,喜好，关注的人还有他的朋友(既如下那些用户)。图1显示了一个典型的Pinterest用户配置文件。就像Twitter,Pinterest的用户在登陆时会出现一个按时间戳排序的时间线。不像Twitter,Pinterest的用户也能关注别人，或者关注一个特别的板块。在前一种情况里，所有朋友的pins都会出现在时间线上。在后一种情况里，只显示从这个特别板块里获取的内容。

**2.2 Twitter**

在大多数人的眼中，Twitter就是微博的标准。Twitter成立于2006年，拥有6.45亿用户，且每天有超过500万条140字的微博。Twitter是目前流行程度第11名的网站。在Twitter上，用户创建一个个人的概述，其中包括概述图，一个简短的自我介绍，其他用户发表的微博，他的粉丝还有他的朋友。每一条微博都有相应的时间标记，并可以发布图像，网址，感情标签，还有地理位置信息。用户可以转发别人的微博给自己的粉丝，虽然Twitter 允许用户将自己的个人信息设为受保护（在这种情况下，只有有权限的用户能够查看他的资料），绝大多数的用户是默认的情况，所有的数据都是公开的 (Wasserman 2012)。

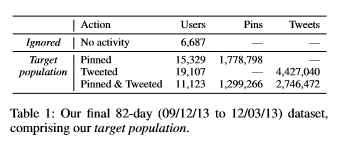
**3.数据和相关方法**

在本节中，我们将分析我们收集到的数据。首先，我们要介绍原始的数据集，并说明是如何收集到他们的。其次，我们对收集到的Twitter和Pinterest数据进行简单的比较，以便去量化目标中的偏差。第三，我们描述的方法用于标记每一条推的类别(比如说，设计，时尚，政治)。这种分类有助于我们对Pinterest和Twitter进行比较。因为所有的pin都是默认有分类的。最后，我们将要介绍如何用用户的活动时序来分组tweets和pins。

**3.1数据和数据收集方法**

我们的目标是比较所有Pinterest和Twitter用户的行为。为了进行这种比较，我们需要找到同时有这两个社交网络账号的用户。幸运的是，有一个简单的方法来识别这些用户：在Pinterest注册用户时，创建一个帐户的唯一方法是登录Facebook或Twitter帐户。 8月间21thandOctober9th，2012年，wecrawled2millionusersfrom Pinterest的，其中210426已经签署，与他们的Twitter帐户（Ottoni等，2013年）。这些用户中，76％的受保护或关闭了他们的Twitter帐户，留下了50702用户对于我们的研究。为清楚起见，我们将参照这些用户作为选择的人群。在2012年8月21日到10月9日这段时间内，我们爬取了200万Pinterest的用户，其中的210426个用户是用Twitter进行绑定的(Ottoni et al. 2013)。在这些用户中，76%的用户要么将他们的资料设为隐私，要么已经关闭了他的Twitter账号，留下了50702个用户供我们研究。为了清楚起见，我们以这部分用户作为参考来分析用户的行为。我们需要爬取他们的pins和tweets。为了爬取tweets，我们可以简单的是用Twitter REST API v1.1版本。但是，在爬取pin时，需要解决两个技术上的难题，1）：每一条pin都必须在发布后24小时内爬取，2)：每条pin都必须单独去爬取。24小时爬取的要求来源于这样一个事实，Pinterest显示的是一个相对时间戳，即一个小时前发布的pin显示的是“发表于1小时前”，而过了一天后会显示“发布于一天前”，我们的研究需要高精度的时间，所以，必须在时间戳显示为小时时抓取。第二个挑战来自于Pinterest没有一个公共的API去获取Pins和时间戳。

为了解决这些问题，我们构建了一个分布式的爬虫来爬取Pintrest的数据。Pintreset的网站是一个复杂的JavaScript应用程序，所以我们利用PhantomJS来抓取。PhantomJS是一个可以自动解析JavaScript的简略的Webkit浏览器引擎。我们的爬取工具从每个用户的每日活动记录中抓取pins。不幸的是，这不可能做到，因为Pinterest对每日访问的数量限制小于50K 个用户。要解决这个问题，我们每天从这50K个用户中随机的抽取30K个用户来抓取。



**最终数据** 我们的数据集包括了略少于三个月的Twitter和Pinterest的用户数据。

我们积极爬取了2013 年9月12日到10月3日之间的Pinterest的数据。因为我们有准确的pin的时间戳，我们通过对时间进行限制爬取了这段时间内的tweets。表1显示了这段时间内抽样用户的行为（如pinning或是tweeting）。我们了解到有23313个用户至少在一个网站上很活跃，我们将这23K个用户作为我们剩下研究的观察对象。

**3.2数据集验证**

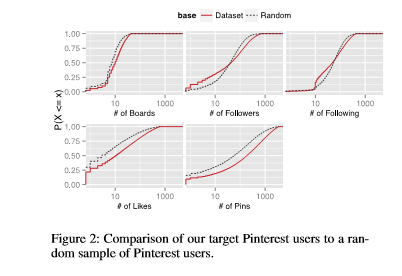
现在，我们开始对Pintreset和Twitter用户的随机样本进行对比。这种比较目标是量化出偏差。为了进行比较，我们均匀地随机地从原来200万Pinterest用户样本中选择30K 用户。同样，我们用同样的方法选择了30K Twitter用户（10K分别来自九月，十月，和2013年11月）。

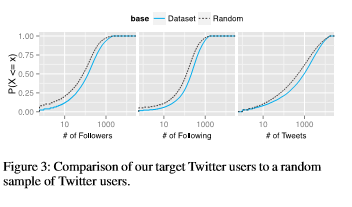
图2比较了我们的Pinterest目标用户和随机用户的特性。目标用户是稍微活跃：他们有更多的板块，喜欢和pins. 这表明，目标用户在社交网络更活跃，这并不奇怪，因为目标用户同时拥有Pinterest和Twitter账户。尽管如此，这两个群体的朋友和粉丝有着很相似的平均数。

图3比较了Twitter中目标用户和随机用户的特点。就像Pinterest，我们观察到目标用户发了比随机用户更多的推。在这种情况下，目标用户也有更多的朋友和粉丝，进一步确认了在社交网络中，目标用户更加活跃。

总体上，图2和3表明存在目标用户和随机用户之间的微小差别。然而，尽管区别很小，我们还是认为，目标人群是具有广泛代表性的活跃的Pinterest和Twitter用户群体。

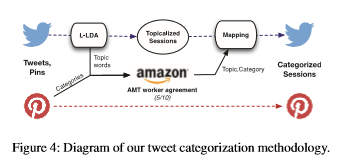
**3.3 Tweets标签的分类**

我们方法的下一步对我们目标用户的tweet进行分类。这一步是必要的，以便使引pin和tweet有直接可比性。

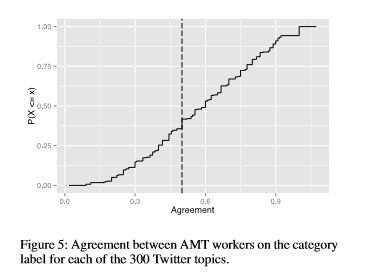
**选择类别** 第一步是在我们的类别中选择相关的类别，我们使用Pinterest提供的33个类别作为一种自然的出发点。并且增加如下五个类别，在对Twitter的相关研究中，他们已经被广泛的定义和接受(Duh et al. 2011; Hong and Davison 2010; Lee et al. 2011; Zhao et al. 2011): “政治学”，“宗教”，“慈善”，“Twitter特有”和“商务”。“Twitter特有”是指讨论Twitter的文化，例如，#followfriday。

为了确定这些类别是否足够，我们用Amazon Mechanical Turk (AMT)来对推特进行上面提到的38个类别进行分类，这项实验揭示了一些关于分类的问题。例如，AMT很难区分”男性时尚“与”女性时尚“，所以我们把这两个类别合并为”时尚“。同样，我们合并了“健康与健身”和“体育”，“家居装饰类”和“DIY”。我们还合并了“孩子”和“儿童与家庭”。经过这些过程，我们留下了35类。、

**Twitter上确定的主题** 第二部是从我们的Twitter语料库中提取主题。这里的关键是用一个知名的主题提取工具来实现分类。如果我们可以从Twitter中提取主题，并映射到我们的35个类别，那么我们也就可以用类别来分类twitter就像用主题分类一样。

为了从我们的tweets中提取主题，我们利用Labeled Latent Dirichlet Allocation (L-LDA) (Ramage et al. 2009)进行提取。L-LDA使用了与LDA相同的底层机制，但是每一个主题都带有一个有标签的种子（即，由研究者选择一个字），以用来对提取进行帮助。我们选择了L型的LDA算法因为以前的工作已经表明，它比LDA能更有效地从微博中提取的主题(Quercia, Askham, and Crowcroft 2012; Ramage, Dumais, and Liebling 2010)。

我们用L-LDA算法对我们从语料库里随机抽取的200万微博数据进行了处理。这是我们在一台内存为70GB的机器上能处理的最大微博数量。我们预先去掉了300个词语（停用词），网址，还有出现次数小于10次的词语。如图4所示，我们设置了LLDA的算法参数如下，α=0.167，β= 0.001，k=300。我们选择了300个最常见的哈希标签作为我们主题的标签。L-LDA算法的最终输出为300个主题，每一个主题都大约包含了50个词，他们由主题的频率排序。

**把主题映射到类别** 我们分类处理的第三步是把从微博中提取的主题映射到35个分类上，如图4所示，我们用AMT执行此任务，我们让每一个Turker用L-LDA算法提取20个主题，并让他们为每个主题选择三个最适合的分类。每一个AMT都被分配了25个与主题最相关的词语。我们把主题分为15组，并且用10台AMT完成每一个任务。AMT的收费是$0.25/台/任务，这需要两天的时间来完成。

从AMT接受到结果后，我们分析数据，以确定由L-LDA算法得出的300个主题能很好的映射到35个类上。图5显示出了每一个主题被AMT映射到分类的认同度。无论他们是被给予了什么类别去区分。 60%的主题在10台AMT上有超过50%的认同度，这意味着我们可以高度相信这是最好的主题和类别之间的映射。在剩下的我们的分类方法中，我们只使用认同度大于50%的主题。我们放弃了低认同度的主题。

**Twitter分类** 我们分类处理的第四步是用推特的主题描述来计算最它可能的分类。最初，我们试图使用LDA算法来为每一条推产生一个主题分布（LDA的输出是一个主题分布，表示一个给定的文件中包含的每个主题的可能性）。然而，我们同个三个独立的评分程序来主题分布的结果，发现主题标签的准确率约为30%。这里低精度的事实来源于每一条推都很短，并且他们没有相应的上下文解释。

为了解决这个问题，我们把推分组成为会话。我们定义一个会话就是一个用户一系列的推。其中，任意两条推之间的时间间隔小于2小时。直观的看，连续的短时间的推应该是相同的主题。所以我们选择2小时作为会话之间的分隔单位。因为2小时是最小单位，所以我们可以准确的选取Pinterest的数据(回忆一下，Pinterest的时间戳是相对的，如”发表于4小时前”)。在3.4节，我们也将把pins分组到会话因此，我们可以直接使用Pinterest和Twitter的数据进行比较。

现在，我们可以为每段会话生成一个主题。我们把每个会话作为一个文档用LDA算法产生主题分布。AMT产生的结果给了我们主题和类别之间的映射，然而，这是一多到一的映射（即，很多主题可以映射到同一类别）。因此，要确定一个主题的范畴，我们从每一个主题概率中减去1/主题数（即基准线概率）。然后计算所有主题到每一个类别的概率和。我们把每个会话的主题指定为概率和最大的那个类别。

最后一步，我们人工的验证了100个随机分配的会话。和以前一样，三个独立的评分者检查每一个会话，以确定是否已被分配的类别适配。35个类中，有11个类别被判定为具有60%以上的精准度（即，大多数情况下，会话的类别准确描述了他的类别）。这11个类别分别是：“艺术”，“慈善”，“设计”，“DIY工艺品”，“时尚”，“食品及饮料”，“头发和美丽”，“健康，健身和体育”，”政治”，“技术”和“婚礼”。由此，我们在第四节中，将重点关注这11类Twitter的会话，而忽略其他类别的会话。

我们把分类Twitter的代码和数据放在研究交流的话题下。

**3.4 Pinterest和Twitter里的会话**

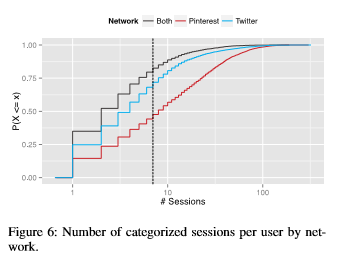
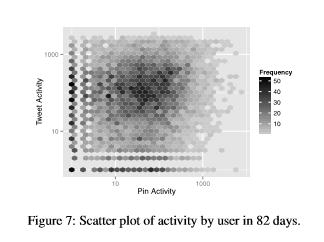
关于这点，我们已经把我们目标用户的推分组为会话并且进行了相应的分类。在过滤了低精度的类别后，我们还剩下100,212个会话（排除了665,980个会话）。为了确保我们的Twitter数据和Pinterest数据能进行比较，我们也把Pinterest数据按2小时为单位分组成为会话。类别为一段会话中出现最多的Pinterest类别。

图6显示了数据集中每一个用户分类的会话数量。我们的用户往往是Twitter的类别比Pinterest的少。这是可以理解的，因为我们已经过滤了低精度的Twitter类别。图6中最重要的数据是“Both”,它绘制了用户在两个社交网络上至少为k的类别。由于我们的目标是比较用户在Pinterest和Twitter上的行为。我们把重点放在同时活跃在两个网络上的用户。由此，我们在第四节中，我们把注意力放在同时在两个社交网络上拥有超过7个话题的用户。总体而言，这个数据集包括1227个用户产生的41928个Pinterest会话和33452个Twitter会话。

**4分析**

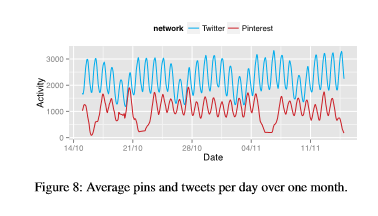
现在，我们开始对在Twitter和Pinterest上跨用户的行为分析。我们首先探讨两个网站的用户的总体模式。（4.1节），随后是对整个网站上用户 的相似性进行实验(4.2节)。然后我们基于上一节的分类调查两个站点间的相同内容。我们首先找出总体来说流行的类别（4.3节），其次是跨站点转发的内容(4.4节)。

**4.1总体分析**

我们通过Pinterest和Twitter用户高层次的特征开始我们的分析，在本节中，我们2万3千位用户82天的1.8M pins 还有4.4M tweets。

首先，我们考察Pinterest和Twitter的整体行为。图7是显示了用户pins和tweets数量的点图。黑色代表大约50个用户，灰色区域代表小于10个用户。

图7显示大多数我们的目标用户都落在轴上下的两个区域之一。有一组同时活跃在Pinterest和Twitter的用户沿着轴分布。相反，有一组用户以100条tweets和100条pins的区域为中心分布。这个结果表明我们的目标用户包括了许多同时活跃在两个网站的用户。



接下来，我们研究Pinterest和Twitter用户的时间动态。图8展示了2013年10月15日至11月15日间用户发的pins和tweets的平均数的变化。从图8中能够得出两个结论：第一，我们的用户的tweets大约是pins的两倍。第二，Pinterest的和Twitter表现出不同的时间动态。Twitter在周一至周五是使用高峰，而Pinterest的高峰在周末。有一个偏差发生在10月28日，我们猜测是因为万圣节原因。

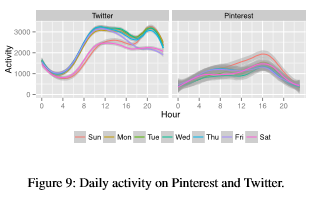


图9通过对用户每周每天不同时段的动态描绘了我们目标用户的日常活动模式。我们的目标用户主要分布在美国，图9是将时间归一化到东部标准时区的结果。显然，无论哪个社交网络的用户，清晨都是他们的使用低谷。然而，Twitter用户经常在深夜发推，并且在周六周日有明显的下降。与此相反，Pinterest用户一直呈缓慢上升的趋势，在星期日到达最高峰。

**4.2 用户语言特性**

在本节中，我们研究了23000个用户的个人描述和他们语言上表现出的特点。我们提出一个问题，用户在Pinterest和Twitter上有着独特的身份还是他们在两个社交网络都共享着同样的身份。为了量化用户的身份，我们考察了三个文本特征：用户名，个人信息的描述，还有在tweets, pins上的用词。我们将注意力放在文本分析上是因为以前的研究成果已经能够成功的对文字表现出来个人情感和认知进行自动分析(Tausczik and Pennebaker 2009; Kahn et al. 2007; Veltman 2006)。

首先，我们比较了用户在Twitter和Pinterest上个人用户名的选择。有研究表明，对于在线系统中个人名称的选择反映着一个人的性格。反映着他们希望被别人认同的方式和他们的文化背景(Suler2002; Bechar-Israeli 1995)。因此，我们推测，如果用户在两个社交网络上使用了相同的名称，说明他们在两个网络上有相同的身份。

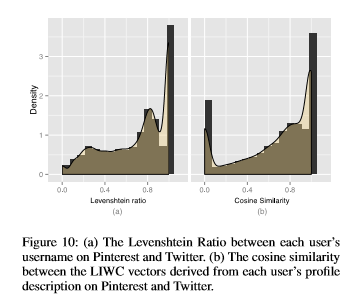


图10（a）绘出了每一个用户不同用户名的莱温斯坦率(LR)。LR的定义是LR(si,sj) = 1 − ED(si,sj)/max(|si|,|sj|)，其中si和sj是字符串，ED（si,sj）是莱温斯坦率中定义的si和sj的距离，|s|是字符串s的长度。简单来说，ED(si,sj) = 1代表两个字符串完全相同，而ED(si,sj) = 0代表他们完全不同。如图10（a）所示，我们的目标用户名称有很高的比例是相同的或是相似的，这表明，用户在两个网络间共享着同样的身份。

其次，我们比较了他们在个人资料里的描述。我们用LIWC(Pennebaker, Francis, and Booth 1999; Pennebakeretal.2007)分析了每一个描述，产生一个词性的向量。我们在图10（b）中绘制了用户Pinterest和Twitter数据LIWC分析结果的对比，就像在图10（a）绘制的那样。和图10（a）得出的结果一样，广大的目标用户在不同网络上有着很大程度的相似度。这一点再度证明了我们选择的目标用户在Twitter和Pinterest有着同样的身份。然而，不同于图10（a）的是，有一部分用户的相似度为0，这表明这部分用户出于不同的目的在使用Pinterest和Twitter,或者是他们希望把他们在这个网络间的身份分开。

第三，我们比较了他们在Tweets和pins上的用语，尽管Pinterest是基于图像的社交网络，但是用户可能会写一个简短的描述。这引出了一个问题，用户是否在跨平台分享时使用了相同的语言风格？要回答这个问题，我们把一个用户所有的关于pin的描述都放入一个单独的文档中，并用LIWC分析了文本内容。对用户的tweetes我们也进行了同样的处理。这将产生一个Pin的向量和一个tweet的向量。为了比较每一个用户向量中的常用语言。我们通过随机的在pins和tweets里抽样并将他们串联起来创建了一个网络版的LIWC向量。并对这个LIWC向量进行分析。

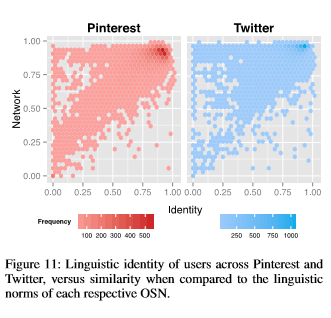
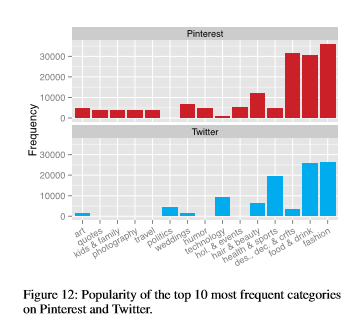


图11给出了对tweets和pins的LIWC分析结果，两个x轴有着相同的样本数量，每个用户的Pinterest和Twitter的LIWC相似性我们成为“身份”轴。Pinterest的y轴显示每个用户的LIWC结果和关联的LIWC结果的相似性，Twitter的y轴类似。我们称之为“关联”轴。比如，一个在（0，0）这个位置的用户，他的Pinterest和Tweeter的相似度为0，以及他们的pins和其他用户Pins的相似度也为0。与此相反，在位置（1，1）的用户，他的Pinterest和Tweeter的完全相似，他和其他用户的相似度也为1。多色图11中给出了一定用户数的相似度结果图。

下面是几个从图11得到的结论。首先，大多数的用户集中在图的右上角，这意味这这些用户在两个网站上有很高的相似度，而且他们和同网站的其他用户也有很高的相似度。这表明，这两个社交网络有着相似的用语趋势，并且大多数用户追从这种趋势。第二，有许多用户集中在左上方而不是右下方，这些用户倾向于使用Pinterest和Twitter上不同的语言趋势，然而，这些用户还是跟随着着平台的语言趋势。这表明虽然一个用户的语言习惯常表现出跨平台性，但还是倾向于采用每个平台的趋势。此前的研究已经发现了网络上类似的语言适应(Danescu-Niculescu-Mizil et al. 2013)。

**4.3 Pinterest和Twitter类别**

在本节中，我们分析Pinterest和Twitter的类别特性。为了便于分析，我们利用3.4节中已经分好类的41928条Pinterest和 33452条Twitter。为了清楚起见，我们围绕本节四个问题来组织。



我们提出的第一个问题非常简单：在Twitter和Pinterest上哪一个类别是最流行的？为了回答这个问题，图12绘出了在这些会话中，Twitter和Pinterest上分别最流行的10个类别。其中两个类别（“时尚”和“餐饮”）是两个站点都最流行的，而“设计，装饰品，工艺品”和“美容美发”在Pinterest上更受欢迎，“健康与体育”则在Twitter上更加流行。虽然其他类别，比如说“技术”和“政治”在Twitter上非常流行，而我们的用户并不喜欢。但是不足为奇，因为我们的目标用户包含了许多Pinterest的喜好者，Pinterest上流行的是设计，食品和美容。

我们要要问的第二个问题是：用户是如何分配他们在不同网站的时间的?用户是否高度的聚拢（即他们的关注内容大部分是哪几类），或者是用户很多样化？为了回答这个问题，我们通过用户的会话来计算他们在不同类别中的香农熵，香农熵的定义如下：

其中，p是用户会话的概率，n是用户会话的数量，是将给定会话映射到相应类别的概率。只关注一类话题的用户熵为0，关注分布很均匀的用户有最大的熵。

图13（a）显示了我们目标用户的会话熵的CDF分布。图13（a）显示，尽管Pinterest有着更广泛的流行类（图12），但是Pinterest用户往往比Twitter用户更加专注。这个结果似乎表明，与Twitter做为一个一般意义上的沟通平台不一样的是，Pinterest则是围绕这特定的兴趣组（如时装，美食，摄影）来发展的。

我们要问的第三个问题来源于第二个问题:我们能不能用户的Pinterest关注类别来预测出他的Twitter关注分类?为了解决这个问题，我们使用关联熵，他的定义为:

其中，p和m是用户Twitter和Pinterest会话的分布率，关联熵用来比较两个概率分布，如果p = m 则 H(p) = H(p,m),因此，越接近的H(p,m)是H(p)的真正熵（例如，1），最好的情况是m是p的近似。简单来说，关联熵接近0的话，则m是p的一个很好近似。

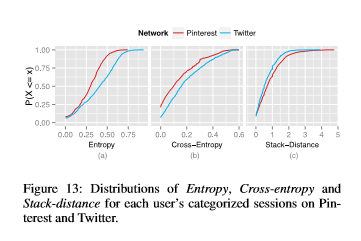


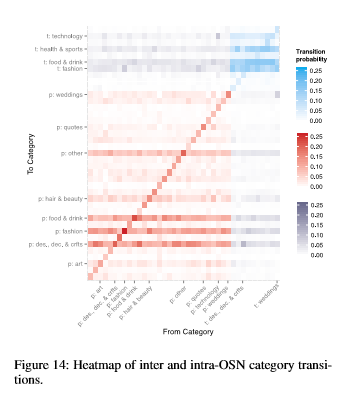
图13（b）画出了我们目标用户关联熵的CDF分布。红线代表H(Pinterest,Twitter)(即，Twitter是模型分布)，而蓝线代表H（Twitetr,Pinterest）。图13（b）显示我们用户的Twitter关注类别能够很好的预测Pinterest关注的类别。这一发现可能源于这样一个事实，即Twitter上流行的分类比Pinterest的要多(见图12)，从而使Twitter的类别更加具有预测能力。

最后一个问题是：是否用户的每一个会话都展现了他的兴趣?具有较高地区特征的用户在连续的会话中展现出了同样的兴趣，较低地区特征的用户出现了随机的兴趣特征。为了对用户的兴趣进行量化，我们使用栈距离(Cascaval and Padua 2003; Almeida et al. 1996) 来分析每个用户会话的时间线。一个长度为0的堆栈距离表示，每一个会话的分类都接近于用户的时间线。

图13（c）画出了目标用户的堆栈距离分布，在两个网站间显示出了类似的趋势。60%以上的用户的堆栈距离小于1。这表明，两个网站之间的会话都有着相同类别的后续会话。这一发现表明不仅仅用户的兴趣是跨平台的。而且他们的兴趣点长期以来都趋于稳定。

**4.4转发的分类**

在本节中，我们研究从Pinterest到Twitter转发的情况，反之，研究从Twitter到Pinterest转发的情况。具体来说，有这样一个问题:新想法是从哪里产生的：Pinterest还是Twitter。这个问题的本质问题是信息在社交网络中传输的界限。为了解答这个问题，我们将注意力再次集中在3.4节的结果中。



为了着手解决这个问题，图14是我们目标用户的兴趣类别的转换图。我们把他们所有的Pinterest和Twitter的会话变成一个。将他们按时间排序后，观察他们的兴趣类别和社交网络间的变化。

图14的红色区域表示的是Pinterest的变化，蓝色区域是Twitter的变化,灰色是社交网络的变化。我们观察到较亮的区域变化更频繁。变化是从列到行的，例如，如果用户当前的会话是在Pinterest上关于技术的，那么他的下一个在Twitter上的会话为技术的概率为15%。

图14中有三个值得注意的点，首先，存在两个网络间相同兴趣的映射（如类别C到类别C），如热点图上的对角线。这再次印证了堆栈距离的结果（图13（c））。其次，我们观察到所有类别都转换成了Pinterest和Twitter上的流行类，这表现为热点图的水平带。虽然从一个网站到另一个网站的类别转换有发生，但是少于同一网站的转换。

第三十特点网络间的自转换发生频率高于其他。例子包括Twitter的“设计，装饰品，工艺品”转换为Pinterest的“婚礼”，Pinterest的“时尚”和“技术”转换到Twitter。这些案列是转发情况或着说信息传递的一个缩影。一个从图14反映出的问题是:为什么不同网络间的转换没有网络自己内部的转换频繁?在今后的工作中，我们计划研究自多任务的apps比桌面任务更多以来，移动社交应用在这个过程中所扮演的角色。

**网络倾斜** 图14显示了网站内的兴趣转换是更为常见的，他并没有告诉我们信息传递的方向，哪一个社交网络是转换的来源。一个新的观点是来源于Pinterest还是Twitter,或者是反过来的？