切尔西赢了，你买了一件t恤:描述Twitter和电子商务之间的相互作用

张海鹏* 1,Nish Parikh2, Gyanit sing2, Neel Sundaresan2
1印第安纳大学信息与计算学院，布卢明顿，IN 2eBay研究实验室，圣何塞，CA, USA
邮箱:zhanhaip@indiana.edu， *{nparikh,gysingh,nsundaresan}@ebay.com*

*摘要:Twitter*和Facebook等社交媒体网站的流行为理解社交媒体和电子商务的相互作用提供了有趣的研究机会。直到最近，大多数关于在线行为的研究主要集中在社交媒体行为和电子商务行为上。在我们的研究中，我们选择了一个特定的全球电子商务平台(eBay)和一个特定的全球社交媒体平台(Twitter)。我们量化了这两个单独趋势的特征以及它们之间的相关性。我们提供的证据表明，大约5%的一般eBay查询流与相应的Twitter提及流表现出强烈的正相关性，而趋势eBay查询流的百分比跃升至25%左右。eBay搜索的某些类别，如“视频游戏”和“体育”，更可能有很强的相关性。我们还发现，eBay的趋势滞后于Twitter的相关对和延迟不同类别。我们展示的证据表明，名人在Twitter上的受欢迎程度与他们在eBay上的相关搜索和销售密切相关。相关性和滞后性为未来的应用提供了预测性见解，这些应用可能会为卖家和电子商务平台带来即时销售机会。

I. 介绍

社交媒体的爆炸使得舆论、意图和观察的传播成为可能，有时甚至比传统新闻媒体的传播速度还要快[1]-[3]。这使得许多挖掘(社会)网络的应用程序能够产生政治、科学和现实世界的价值，包括预测选举结果[4]、[5]、监测流感的传播[6]、模拟雪[7]和地震[3]等生态现象。

在理解搜索和电子商务网站上的用户会话、行为和活动方面也有大量的工作。这样的分析可以通过理解用户意图来释放商业价值。最近在这方面的工作包括量化用户的领域知识对其搜索行为的影响[8]，比较用户在不同设备上的搜索行为[9]，理解用户在查询没有返回结果时的行为[10]，并重写这些查询以提高召回率[11]。用户的竞价行为也被研究过[12]-[15]。

我们认为，对社交媒体和电子商务领域的行为特征的理解和它们对彼此的影响将创造具有社会和经济价值的机会。通过分析中国电子商务网站的内部社交网络，我们朝这个方向迈出了一步，模拟了买家将为与可信卖家的交易支付多少钱[16]。Facebook口对口推荐与一种特殊的电子商务形式——日交易之间的相互作用已经被研究过[17]。尽管这两篇论文揭示了社交网络和电子商务之间互动的有趣方面，但反映普通公众兴趣和需求的两个领域趋势之间的相关性仍未得到探索。

在本文中，我们通过分析带有时间戳的推文、电子商务搜索日志和交易数据，在大尺度和细粒度上研究了社交媒体网站(我们选择Twitter)和电子商务网站(我们选择eBay)上的用户行为的相互作用。一个假设和简化的场景是:在Twitter上，你看到人们在谈论你最喜欢的足球俱乐部切尔西的胜利，之后你去eBay搜索并最终购买了一件切尔西的t恤。我们会回答这样的问题:社交媒体上人气的突然飙升与电商行为之间有什么关联?当此类事件发生时，在活动的爆发中，是有一个平台领先还是落后?我们的研究结果表明，大约5%的一般eBay查询流与相应的Twitter提及流具有很强的正相关性，而对于趋势查询[18]，这一比例达到25%左右。来自某些eBay类别(如“体育”)的查询具有更好的相关性。我们还发现了eBay在相关流对方面落后于Twitter的证据。通过监测名人名单的受欢迎程度，我们发现他们的Twitter受欢迎程度与相关的eBay搜索和销售相关。相关性和滞后性对预测任务很有用。例如，通过监控社交媒体，可以立即决定一个爆款是否会带动电商平台的销售，因此，相关卖家和潜在买家将提前得到通知，以刺激交易。

在本文的其余部分中，通过对相关工作的调查，我们描述了量化两个流之间的相关性以及两者之间的滞后的方法，并将它们应用于从数亿条eBay搜索日志和推文中提取的时间序列。我们进行了一个案例研究，在此之后我们结束了我们的工作。

*This work was done while this author was an intern at eBay Research Labs.

II. 相关工作

最近的许多研究都集中在挖掘(社会)网络，以创造实用、科学和社会经济价值。

*事件检测和趋势分析*对传统新闻媒体语料库进行了分析，以探索时间特征。Kotov等[19]从多语言网络新闻流中提取具有相关时间爆发的命名实体。Cook等人[20]在跨越一个世纪的新闻报纸语料库中测量人们的成名期。Radinsky等人[21]用新闻档案中相关术语的时间分布来表示单词语义。也有关于用户生成的网络数据的时间分布模式的研究，包括搜索日志[18]，[22]-[24]，标签[25]，[26]和推文[27]，[28]。由于推文是用户表达自己的想法、观点和兴趣的短文本，它与其他用户生成的网络数据或新闻文章有本质上的区别，因此应该区别对待。在Twitter上的事件检测方面已经做了很多工作。Weng等[29]，Becker等[30]和Sakaki等[3]研究Twitter中的文本流来检测现实生活中的事件。Ritter等[31]不仅检测了Twitter上的事件，还对事件进行了分类。Marcus等人[32]构建了一个系统来可视化和总结Twitter上的事件。与之前的工作不同，我们的论文重点是将Twitter流与全球电子商务平台的相应流进行比较，以研究相关性。

*建模和预测*已经挖掘了各种web语料库，用于其他领域的建模和预测，以创造社会经济价值。Bollen等人[33]通过估算Twitter上的公众情绪来预测股市指数，但并未探索社交媒体与商业直接相互作用的其他可能性。Jin等[4]通过监测Flickr照片来估计产品的采用率。由于Flickr具有照片共享的特性，并且与Twitter相比，它的用户基数相对较小，因此仅对iPod等现象级产品进行了粗时间粒度的估计。Choi等人[34]和Goel等人[35]使用搜索引擎查询日志来预测视频游戏和机动车辆的销售、电影票房收入和失业率等事实。虽然大量的查询反映了普通大众的意图，但他们并没有在用户明确表达想法的社交媒体上与可以即时监控大量交易的全球电子商务平台上对用户行为进行并行和即时的比较，因为票房收入等事实通常是延迟发布的。这些研究的结果揭示了我们试图理解另一种联系的尝试，这种联系是在一个流行的社交媒体网站和一个全球电子商务市场之间，在大规模和精细的层面上进行的。

为了为用户提供更好的服务，增加网站的利润，我们分析了电子商务网站的用户行为。很多工作都是基于对搜索查询日志的理解。Parikh和Sundaresan[18]开发了一种近乎实时的突发检测系统，为买家和卖家提供趋势查询建议。Singh等人[10][11]研究了用户在初始查询没有返回匹配结果的情况下的行为，并构建了一个系统来重写这些查询以提高召回率。关于在线拍卖的用户行为一直是研究的另一个主题，包括交叉竞价[14]、最后一刻竞价[12]、先令[13]和卖家声誉对价格的影响[15]。除了只关注电子商务领域本身，我们还研究了社交媒体领域，以量化相关性并探索是什么推动了电子商务网站的销售。

*社交媒体与电子商务*的交叉研究带来了有趣的发现。Guo等人[16]分析了中国电子商务网站上买家和卖家的内部社交网络，以模拟买家会为与可信卖家的交易支付多少钱。重点是内部社交网络的结构和信任的价格，而不是像我们在本文中那样分析开放社交媒体平台和全球电商网站上的趋势和用户行为。Byers等[17]研究了团购网站(包括Groupon)及其与Facebook、Yelp等社交媒体网站的关系。在他们的研究的非社交方面，他们分析了用户购买的动机，并使用交易的参数来预测销售额。在社交方面，他们研究了日常交易如何影响商家在Yelp上的声誉，他们认为Facebook上的口碑推荐有利于日常交易网站。他们研究的交易是电子商务的一种特殊形式，在相对较短的时间内为消费者提供本地化的折扣商品，而我们则在大规模和长期的两个领域中对更一般的用户行为进行研究。此外，我们没有研究社交媒体中的信息扩散，而是专注于分析信息的时间属性，这些属性反映了用户的即时兴趣和需求。

III. 量化相关性

我们从两个数据集开始——Twitter Tweet和eBay搜索查询日志。Twitter数据集包含一个Tweet示例，每个Tweet都有Tweet本身的简短文本、组成Tweet的人的用户id和指示Tweet发布时间的时间戳。eBay数据集包含eBay搜索查询日志的一个样本，每个日志都有一个文本查询、发出查询的人的用户id、用户点击的结果项目的数量、用户收到的结果项目的出价数量、用户购买的固定价格“Buy It Now”(BIN)项目的总数以及用户为这些BIN项目支付的总价。

我们专注于出现在Tweet和eBay查询中的关键字短语，对于一个关键字短语，我们提取了其提及的两个时间序列，分别作为其在Twitter和eBay上受欢迎程度的代表。对于一个关键字短语，如“巴拉克·奥巴马”，在一定时期内，在每个时间单位，我们计算包含这个关键字短语的Tweet的唯一Twitter用户的数量，以及查询包含这个关键字短语的唯一eBay用户(“SEARCH”)的数量。对于eBay数据集，除了每个时间单位的唯一用户数量外，我们还可以计算结果物品的总点击次数(' VIEW ')和出价(' BID ')，结果物品的购买BIN数量(' BIN计数')，这些物品的总BIN数量(' BIN总数')和平均BIN价格(' BIN平均价格')。

如[36]所示，我们计算时间序列的归一化对之间的Pearson Correlation系数为

以及用于检验真实相关系数等于0的零假设的双尾p值。在下面的小节中，我们将详细描述时间序列提取和相关计算。

A. 时间序列提取

我们从两个数据集中提取代表关键字趋势的时间序列。

一个使用eBay的买家的简化场景是:用户发出查询并点击几个结果项目，然后用户可以选择对其中一些项目出价，她也可以选择使用“Buy It Now”(BIN)立即购买一些固定价格的项目。在实验中，我们定义了所有用户动作的集合 *A = {a1, a2, ⋯, aq}* 作为 *ai* = (ui, *qi, ti, vi, bidi, bini, pi) 形式的元组集合*，其中*ui*是用户，*qi*是查询，*ti*是时间戳。*vi*是用户发出查询后在结果项中唯一点击的项数，*bidi*是用户对结果项的出价数，*bini*是用户在结果项中购买的BIN项数，*pi*是用户为购买的所有BIN项支付的总金额。

为了得到关键字短语*k*从时间ts到时间*ts*在eBay上的搜索提及的时间序列，我们将这段时间分为*m*个粗时间bin。对于每个bin，我们计算发出包含关键字短语的查询的唯一用户的数量。设*quant(ti, ts, te, o)*是一个量化函数，它将时间戳*ti*映射到*m*个时间bin中的一个，时间bin的起始时间为*ts*，结束时间为*te*，时间bin的大小为*o*，返回一个在[1,*m]范围内的bin索引*。设*ngram(qi)*是一个函数，它将文本字符串*qi*中的非数字和非字母字符替换为空格，并返回新字符串中所有可能的*ngram*的集合。对于任意关键字*k*，我们然后构建一个*m维*向量，计算在从*ts*到*te*的每个时间周期*b*中，发出*n*-gram包含*k*的搜索查询的唯一用户的数量。

$$U_E(b,k,t_s,t_e) = ||\{u_i|(u_i,q_i,t_i,v_i,bid_i,bin_i,p_i) \in A, \\ k \in ngram(q_i), b = quant(t_i,t_s,t_e,o)\}||,$$

其中*A*包含了从*ts*到*te的所有A*。使用这种方法，我们还可以提取' VIEW '，' BID '，' BIN计数'，' BIN总数'和' BIN平均'的时间序列，如上所述。

类似地，我们提取Twitter上关键词短语*k*的提及时间序列。我们构建一个*m维*向量*UT* (b, *k, ts, te)*，计算在从*ts*到*te*的每个时间周期*b*中发布推文的唯一用户的数量，其中*n*-gram包含*k*。

对于提取的所有向量，我们执行*l1范数*。

B. 皮尔逊相关系数和t检验

我们计算每个关键词短语的eBay向量*E*和相应的Twitter向量*T*之间的皮尔逊相关系数*r*。它测量了两个数据集之间的线性相关性，并给出了[−1,1]中的协效率，其中+1表示精确的正线性关系，-1表示精确的负线性关系，0表示没有相关性。它的计算方法是*E*和*T*的协方差除以它们的标准差的乘积:

$$r = P(E,T) = \frac{cov(E,T)}{\sigma_E \sigma_T}$$

我们对这对向量的皮尔逊*r*进行学生*t*检验，零假设是它们不相关(实际*r*为0)。如果潜在变量遵循二元正态分布，则皮尔逊*r*的抽样分布将具有自由度*n* = 2的学生t分布。根据[37]，这大约适用于不小的样本量，即使观察值不是正态分布。因此，即使*E*和*T*可能没有二元正态分布，我们仍然可以使用T检验中的双尾p值作为相关性置信度的近似值。

IV. 计算延迟

通过应用移动窗口方法，我们量化了一个流在另一个域中滞后于其对应部分的证据。它找到了一个位移，使这对流之间的皮尔逊*r*最大化。虽然来自一对流的证据并不强，但当我们计算大量关键字短语的滞后时，就会有足够的信号提供见解。我们选择这种方法，而不是[33]中建议的流行的格兰杰因果检验[38]，因为固定的时间箱不适合捕捉和量化流行度的瞬间突然变化。例如，当bin大小固定为一天时，如果一个流滞后于另一个流几分钟或几小时，它可能无法捕获一天内发生的变化;如果将bin大小设置为几分钟或几小时，则每个bin中可能没有足够的计数;对于不同的流对和不同类型的事件，延迟也可能会有所不同，这些事件可能只会被以更细粒度移动的Windows捕获。

对于每一对流，我们将一个流移动一段时间Δt以计算向量，并计算移位向量与未移位的另一个向量之间的皮尔逊相关性。我们定义位移的时间段Δt，它使皮尔逊的*r*最大化作为滞后。我们在这里给出正式的定义。给定*k*、*b*、*ts*、*te*, eBay向量与Twitter向量的皮尔逊相关系数计算为:

$$f(\Delta t) = \begin{cases} P(U_E(t_s+\Delta t, t_e), U_T(t_s, t_e-\Delta t)), \Delta t >= 0 \\ P(U_E(t_s, t_e-|\Delta t|), U_T(t_s+|\Delta t|, t_e)), \Delta t < 0 \end{cases}$$

其中一个流被Δt平移。*K*和*b*被省略，因为它们在比较中的所有流中都是固定的。如果Δt >= *0,则*eBay流的起点被移到较晚的时间戳，同时Twitter流的终点被移到较早的时间戳，以确保两个结果向量具有相同的维度。如果Δt < 0，则Twitter流的起点将转移到较晚的时间戳，eBay流的终点将转移到较早的时间戳。延迟l计算为:

$$l = \underset{\Delta t \in [-\mathcal{T}, \mathcal{T}], \Delta t \in \mathbb{Z}}{\operatorname{argmax}} f(\Delta t),$$

其中，正l表示eBay落后Twitter l，而负l表示Twitter落后eBay l。

V. 实验和结果

A. 一般相关性

我们从eBay查询的长尾分布中选择最热门的查询作为我们监控的一般关键字短语。对于eBay(2012年7月11日)上排名前15万的查询，我们监控了eBay和Twitter从2012年1月1日到2012年3月31日这段时间。在本实验中，eBay数据集包含查询日志的大量代表性随机样本，Twitter数据集包含Twitter上发布的1%的tweet。我们为每日查询计数和每日Twitter提及设置阈值，以确保我们提取的向量的密度。对于平均eBay每日查询计数，我们要求数字不少于20，对于平均Twitter提及，我们将阈值设置为每天5。这为我们提供了一个16,099个关键字短语的列表，可以代表eBay上的一般查询。为方便以后的引用，我们将其命名为General QueryList。对于这个列表中的每个短语，我们从eBay和Twitter上获得两个每日被提及的时间序列。对于每一对时间序列，我们计算Pearson相关系数r和t检验的p值，假设两个时间序列是由两个不相关的系统产生的。

表1中的统计数据表明，大约5.25%的这些查询可以被认为与Twitter上的对应查询具有显著的正相关关系，而只有大约0.65%的查询在p = 0.0005时显示出显著的负相关。在这里，我们考虑了从0.0001到0.01的几个置信水平，并计算了正相关和负相关关键字短语对的分数。注意，当我们提高置信水平时，正相关对和负相关对之间的比率增加，表明可能的噪声被去除。

为了检查不同查询类别在两个领域之间的行为相关性，我们将这些查询分解为36个eBay元类别[1]，一些类别在0.0005的置信水平上具有较高的强相关查询部分。如我们所料，“视频游戏”、“体育俱乐部、卡片和粉丝商店”等类别更有可能推动冠军比赛和电子游戏发布等新闻事件的销售，而“家居和花园”、“陶器和玻璃”等类别则排在前5名(见表2)。在所有的一般查询中，“体育俱乐部，卡片和粉丝商店”占了7.34%，这是一个比较大的比例。

然后，我们从上面提到的GeneralQueryList中选择[18]中描述的系统检测到的趋势查询，得到730条查询，我们将其命名为Trend-QueryList。如表III所示，24.93%的趋势关键词短语对具有非常强的相关性，而一般关键词短语对的相关性为5.25%。在0.01，一个宽松的置信水平下，几乎30%的趋势关键词短语是相关的。在同一水平上，正相关对与

[1] There is 36 meta categories at the top level of the eBay merchandise ontology composed by domain experts. The list is available on www.ebay.com. A query is assigned to a category according to what the majority of users issuing the same query clicked on and purchased historically.

表1 .不同置信度下关键字短语相关对的分数。正相关对和负相关对之间的比值随着置信度的增加而增加，提示可能存在的噪声

是删除。

p-value	pos_corr	neg_corr	pos/neg
0.01	8.52%	2.49%	3.42
0.005	7.34%	1.77%	4.14
0.001	5.75%	0.80%	7.18
0.0005	5.25%	0.65%	8.07
0.0001	4.35%	0.32%	13.59

表二。在置信水平上具有强相关性的查询部分排名前5位的ebay元类别

0.0005.

Category	pos_corr
Video Games	21.28%
DVDs & Movies	14.20%
Entertainment Memorabilia	13.47%
Sports Mem, Cards & Fan Shop	13.45%
Tickets	13.04%

负相关对远高于表1。当我们将这些趋势查询分解为eBay元类别时，我们发现一些类别表现出更多的相关性。如表四所示，69.23%的“体育俱乐部、卡片和粉丝商店”查询在0.0005的置信水平上是强相关的。即使是排名第五的“音乐”类别，其百分比也比一般查询中排名第一的类别高15%。

B. 两种流之间的滞后

从GeneralQueryList中，我们使用Pearson r为0.4的阈值获得了一个强相关的对列表，并产生了690个查询，我们将其命名为related- querylist。对于这些查询，我们计算了[−5000,5000]分钟偏移范围内的滞后，正值表示eBay滞后于Twitter，反之亦然。

为了减少噪声，我们要求皮尔森r的增加至少为5%才能认为存在滞后。然后，我们绘制了滞后的直方图，如图1所示。平均滞后值为290分钟(4.83小时)，690个查询中有61.30%具有正滞后值，而如果滞后值在平均值0附近正态分布，则为50%，这表明对于每天相关的关键字对，eBay落后于Twitter。对于eBay落后于Twitter的配对，

表3。对于趋势关键字对，不同置信度下相关关键字对的分数。

在同一水平下，正相关对与负相关对的比值远高于此

在表i .中

p-value	pos_corr	neg_corr	pos/neg
0.01	29.58%	1.64%	18.03
0.005	28.21%	1.09%	25.88
0.001	25.61%	0.13%	197
0.0005	24.93%	0.13%	191.76
0.0001	22.46%	0%	INF

表iv.趋势关键词对，前5名ebay元

类别按查询的部分排序，在0.0005的置信水平上具有强相关性。

Category	pos_corr
Sports Mem, Cards & Fan Shop	69.23%
Video Games & Movies	64.28%
Cell Phones & PDAs	52.94%
Entertainment Memorabilia	37.50%
Music	36.66%

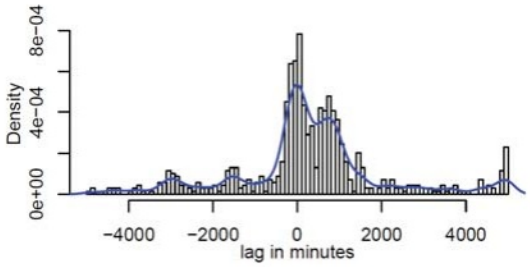

图1所示。相关查询列表的滞后直方图与密度曲线。来自Twitter的eBay的平均滞后时间为4.83小时，690个查询中有61.30%具有正滞后值。大部分滞后分布在x轴的正半部分。

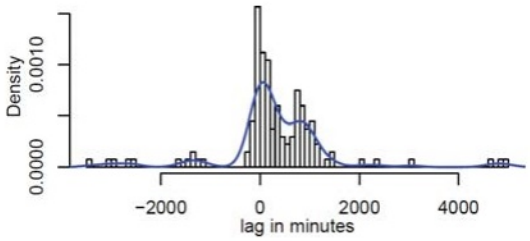

图2所示。相关查询列表中体育滞后的直方图与密度曲线。74%的滞后值为正，因此，大多数滞后分布在x轴的正侧。

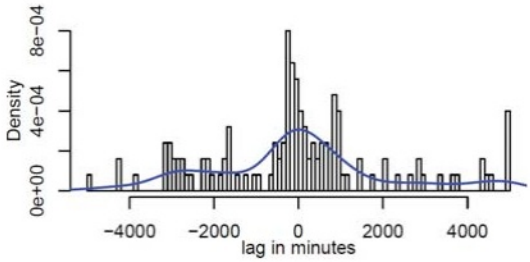

图3所示。相关查询列表中服装滞后与密度曲线的直方图。45.6%的滞后值为正，滞后在0的两侧分布更均匀。

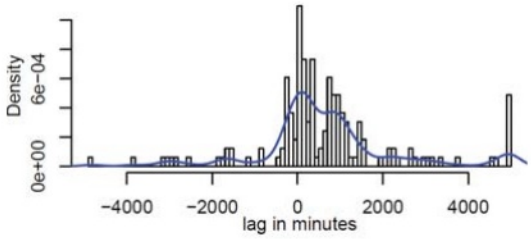

图4所示。趋势关键词的滞后直方图与密度曲线。其中76.82%具有正滞后值，大部分滞后也分布在x轴的正侧。

平均滞后为1214分钟(20.2小时)，第三个分位数为1441分钟(24.0小时)。

在36个eBay元分类中，“服装、鞋子和配饰”(“Clothing”)和“体育用品、卡片和粉丝店”(“体育”)在690个关键词中拥有超过100个关键词，我们选择观察它们的滞后模式，如图2和图3所示。对于“服装”，只有45.6%的人具有正滞后值，而对于“体育”，这一比例为70.14%。这表明，在某些类别中，Twitter领先eBay的信号更强。

再一次，我们从related- querylist中选择趋势查询，它给了我们164个关键字短语。平均滞后值为662分钟(11.03小时)，其中76.82%具有正滞后值，而普通查询为290分钟和61.30%。直方图如图4所示。注意到有10个关键字短语的延迟超过4000分钟，我们在这里用红色的eBay曲线标注了其中两个:一个是一般关键字短语“空调”，延迟4950分钟(3.43天)，另一个是更具体的产品关键字短语“droid 4”，延迟4981分钟(3.45天)，如图5和图6所示。转换后，Twitter和eBay之间关于“空调”的皮尔逊r值从0.42上升到0.48，而对于“droid 4”，该值从0.45上升到0.57，这表明eBay对Twitter这两个关键词短语的响应幅度超过三天。“droid 4”是摩托罗拉在2012年2月10日发布的一款手机，但在此之前3天，有关于Verizon确认发布日期的推特和新闻文章，这导致了2月7日Twitter的爆发。

在图1、图2和图4的直方图中，相邻的局部最大值之间的间隔通常在一天左右(1440分钟)，这可能与用户每天重复的活动模式有关。

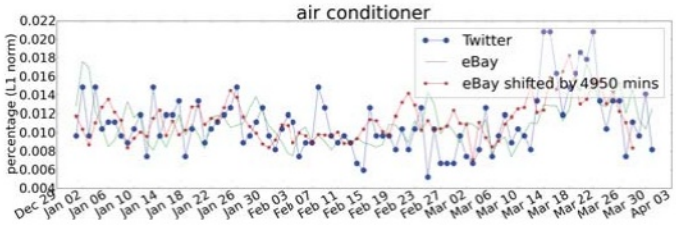

图5所示。Twitter趋势，eBay趋势和eBay的“空调”趋势。滞后(移位)为3.43天。

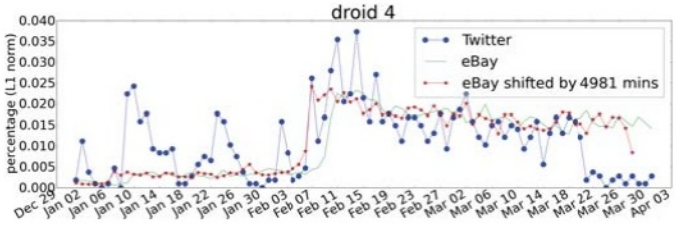

图6所示。Twitter趋势，eBay趋势和“droid 4”转变的eBay趋势。滞后(移位)为3.45天。

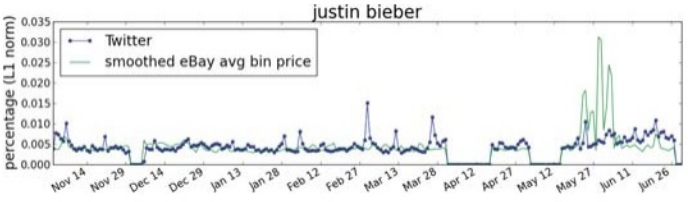

图7所示。Twitter趋势和平滑eBay平均BIN价格的“贾斯汀比伯”，平滑窗口大小为2。与不平滑相比，皮尔逊r从0.183到0.233,p值为0.0007。(零值是由于缺少数据。)

表v对于100个名人关键词，不同置信度下相关关键词对的分数。对于SEARCH和VIEW，6个统计量之间的相关性最好。BID、BIN数和BIN总数适度的相互关系。

	SEARCH		VIEW		BID		BIN_count		BIN_total		BIN_avg	
p-value	p_corr	n_corr	p_corr	n_corr	p_corr	n_corr	p_corr	n_corr	p_corr	n_corr	p_corr	n_corr
0.01	46%	0%	44%	0%	14%	0%	16%	2%	12%	0%	5%	0%
0.005	45%	0%	43%	0%	11%	0%	14%	1%	11%	0%	3%	0%
0.001	39%	0%	38%	0%	10%	0%	11%	1%	9%	0%	1%	0%
0.0005	37%	0%	30%	0%	8%	0%	7%	1%	7%	0%	0%	0%
0.0001	37%	0%	30%	0%	8%	0%	7%	0%	7%	0%	0%	0%

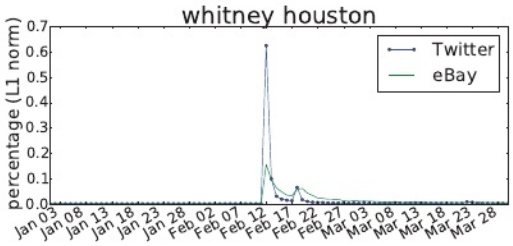

图8所示。“惠特尼·休斯顿”的推特趋势和eBay趋势。在她去世的那天和她的葬礼那天，这两个流媒体上的爆发非常吻合。用户在eBay上的兴趣下降速度较慢，这表明他们的注意力模式与推特不同。

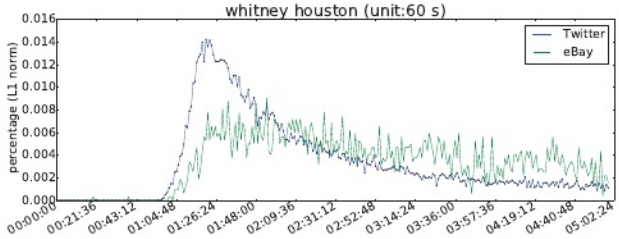

图9所示。推特趋势和eBay趋势为“惠特尼休斯顿”，单位设置为60秒。推特的增长速度比eBay更快，峰值也更早。用户对推特的兴趣在高峰后下降得更快。

C. 名人看

名人新闻经常在推特上成为趋势[1]，我们想知道名人新闻——死亡、电影、被捕、游戏等——何时在一个网络上成为趋势，对另一个网络产生什么影响。一些可能的情况包括:名人死亡的消息引发了一本书的发行，奥斯卡提名之后，名人主演的电影的销量突然增加，名人退休激发了粉丝对纪念品的热情。在这里，我们探讨名人在推特上的受欢迎程度与他们在eBay上的相关搜索和销售之间的关系。我们利用了福布斯的The

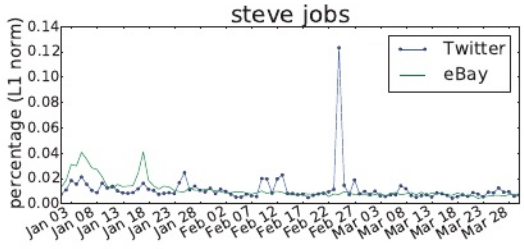

图10所示。“乔布斯”的推特趋势和eBay趋势。eBay上的前两次大爆发与史蒂夫·乔布斯玩偶的生产有关，而推特上的大爆发与他的生日有关。相对较弱的相关性表明，两股流之间的相互作用有时是深远的。

“名人100强”榜单根据媒体曝光率和娱乐相关收入来评选名人。

从2011年11月1日到2012年6月27日(不包括34天)，我们计算每个名人的每日统计数据，包括搜索次数(SEARCH)、观看次数(VIEW)、出价次数(BID)、垃圾箱数量(BIN计数)、BIN总价格(BIN总价格)和BIN平均价格(BIN平均价格)，这样对于名单上的每个名人，将有6个来自eBay的时间序列。对于每个eBay时间序列，我们计算自身与相应推特提及时间序列之间的皮尔逊r，以及t检验的p值。我们在不同的置信水平上设置阈值来检验正相关对的百分比和负相关对的百分比，结果见表V。

结果表明，只有在较低置信水平下的BIN计数才存在负相关对。它还表明，对于SEARCH和VIEW，这对在6个统计量中相关性最好。BID, BIN计数和BIN总数显示出适度的相关性，对此可能的解释是，有时用户只是想查看eBay上有什么，但不一定会下定决心购买。

平均BIN价格(BIN平均值)相关性相对较小，只有5%在0.01的置信水平上相关。一个可能的原因是，价格需要时间来反映受欢迎程度的变化(BIN商品的价格是在卖家列出商品时设定的)，而且价格本质上可能是不敏感的。为了进一步理解这一点，我们将某一天的BIN平均值计算为未来n天的平均BIN价格(BIN总数/BIN计数)，并与当天一起计算。我们在表VI中显示了n在0 ~ 14范围内的平均皮尔逊r，在表VII中显示了$n \in [0,4]$在不同置信水平下相关对的部分。从这两个表中，平均皮尔逊r值在n = 2时达到峰值，同时，大部分(9%)的配对在0.01的置信水平上相关，这表明名人的受欢迎程度对其相关商品在3天内的平均价格的影响。在这里，我们展示了Justin Bieber的一个例子，图7中的窗口大小为2。与不平滑相比，皮尔逊r从0.183到0.233,p值为0.0007。

D. 两条流的峰值(Peakiness)

来自网站的使用流的峰值反映了其用户的注意力以及平台本身的性质。我们通过计算eBay流和推特流的平均秒矩来测量它们的总体峰值，如[26]所示。对于表示时间序列的向量v，其秒矩计算为:

2 www.forbes.com/celebrities/list/, May 2012

TABLE VI. TWITTER提及时间序列和EBAY平均BIN价格时间序列在(n + 1)天窗口之间的平均皮尔逊*r*。当*n* = 2时，皮尔逊平均值*r*达到峰值。

n	0	1	**2**	3	4	5	6	7	8	9	10	11	12	13	14
avg_r	0.018	0.015	**0.020**	0.018	0.012	0.009	0.010	0.013	0.014	0.012	0.012	0.013	0.011	0.007	0.005

表七世。 对于*n*∈*[0,4]*，不同CONFIDENCE下的关键字CORRELATEDPAIRS的部分。WHEN *n* = 2时，大部分(9%)PAIRS在CONFIDENCE为0.01的水平上CORRELATED。

	n=0		n=1		**n=2**		n=3		n=4	
p-value	p_corr	n_corr	p_corr	n_corr	p_corr	n_corr	p_corr	n_corr	p_corr	n_corr
0.01	5%	0	6%	0	**9%**	0	8%	1%	7%	0
0.005	3%	0	4%	0	8%	0	8%	0	7%	0
0.001	1%	0	4%	0	4%	0	4%	0	4%	0
0.0005	0	0	1%	0	3%	0	4%	0	2%	0
0.0001	0	0	1%	0	3%	0	4%	0	2%	0

$$second\_moment(v) = v \cdot v = \sum_{i=1}^{n} v_i^2.$$

对于第V-A节中提到的GeneralQueryList，我们计算他们在eBay和推特上被提及的时间序列的平均秒数。eBay的值是0.011，而推特的值是0.016，这表明推特流比eBay流更尖峰。原因可能是，作为一种新闻媒体，推特用户的注意力上升和下降得更快。

VI. 案例研究

我们选择一些案例来展示我们观察到的两种趋势的特征。“惠特尼·休斯顿”和“史蒂夫·乔布斯”这两个关键词被选中，因为它们既受欢迎，又容易与现实世界的新闻事件联系起来。

图8所示的“惠特尼·休斯顿”流似乎具有很强的相关性(皮尔逊*r* = 0.794, p值= 5.44e-21)。这两条流的爆发在惠特尼·休斯顿去世的那天和她葬礼的那天有很好的关联。但用户的关注模式存在一些差异。作为一个新闻媒体，Twitter更有爆发力，但在爆发力之后，用户获得了信息，就不再谈论它了，而在eBay上，用户仍然对购买感兴趣。

然后我们分析了2012年2月12日00:00到05:00这段时间，在这段时间里，公众开始知道惠特尼·休斯顿去世了。如图9所示，将单位设置为60秒，我们可以看到查询和推文如何达到更细的粒度，并且它表明Twitter的爆发速度比eBay更快，峰值也更早。在这个规模下，我们仍然可以观察到用户在Twitter上的兴趣在高峰后下降得更快，这表明用户在知道消息后仍然有兴趣在eBay上购买。

对于图10所示的“史蒂夫·乔布斯”流，相关性不强(皮尔逊*r* = 0.118, p值= 0.264)。我们观察到1月份eBay有两个高峰，而Twitter上相应的高峰相对较弱，这并不常见，因为Twitter通常比eBay更爆发。我们手动查看了1月3日与eBay第一次大爆发相对应的240条包含“史蒂夫·乔布斯”的推文。其中超过三分之一是关于一家中国工厂计划未经授权制作史蒂夫·乔布斯的逼真复制品[3]，一些推文表示愿意购买。

与此同时，我们在同一天查看包含“史蒂夫·乔布斯”的eBay查询，其中约33%与人物玩偶或摇头娃娃有关，而1月日的比例仅为5%。对于eBay上的第二个突发事件，我们检查1月17日相关的推文和eBay查询。三分之一的推文是关于中国工厂取消玩具公仔的生产，50%的eBay查询与玩具公仔有关。一篇新闻文章[4]指出，人们仍然在eBay上买卖这些公仔，最高价格达到2500美元。

对于2月24日Twitter上的巨大爆发，eBay上并没有明显的相应爆发。但当我们查看当天的推文时，发现那天是史蒂夫·乔布斯的生日，很多人都在推特上纪念他。所有这些都表明，Twitter和eBay上的用户行为之间的相互作用有时可能是深刻的，这种相关性/非相关性可以为发现推动电子商务平台销售的因素提供重要的见解。

VII. 结论和未来的工作

在本文中，我们提出了量化社交媒体趋势与电子商务趋势之间相关性和滞后性的技术。我们还通过案例研究和测量其峰值来检查这两种流的个体特征。我们发现了以下证据:

- 大约5%的eBay查询流与其相应的Twitter流具有很强的正相关性。对于趋势查询，这一比例跃升至25%左右。
- 一般查询的某些类别更有可能具有这样的相关性。例如，对于“视频游戏”类别，21.28%的查询趋势是强相关的，而对于“dvd和电影”，这一比例为14.20%。对于趋势查询，这一比例也更为显著。例如，来自“体育纪念品、卡片和球迷商店”的查询的百分比约为70%。
- 对于相关的流对，eBay流滞后于Twitter流，对于趋势查询和“体育”等类别的查询，滞后更为明显，并且可能对预测任务有用。
- 名人在Twitter上的受欢迎程度与他们在eBay上的搜索和销售趋势有关。还有一些信号

[3] www.foxnews.com/tech/2012/01/03/steve-jobs-action-figure-planned-for-february, [4] www.pcworld.com/article/248238/maker of steve jobs action figure Fox News kills project.html, PCWorld

它们会对相关商品的价格产生影响。

·Twitter的趋势比eBay的趋势更明显。

总而言之，我们观察到电子商务活动与社交媒体密切相关，但滞后于社交媒体，特别是在“体育”和“dvd和电影”等特定领域。当用户对事件和突发事件做出反应时，这一点更为突出。解释这种相关性的一个可能原因是，体育和娱乐领域事件多，吸引眼球，更容易产生具有直接商业价值的新闻。滞后的一个原因可能在于这两个平台的性质，一个是快速生成和传递最新新闻的社交媒体平台，而另一个则满足了对最新新闻的购买需求。我们认为，访问在线社交媒体流可以实现相关产品的近乎实时的销售。

在未来的工作中，我们计划根据社交媒体的信号来预测电商平台的销售情况。最终，该系统将能够检测社交媒体中的事件，将其分类为销售驱动和非销售驱动事件，并估计相应产品在电商平台上的销售情况。有了这样的系统，我们就可以提前很好地为卖家和买家推荐相关的商品，从而增加电商平台上的交易。这将涉及到事件检测、文本挖掘和机器学习等技术。在这个方向上的第一步将是关注查询的子域(例如“体育”和“dvd &电影”)，这些查询显示出很强的相关性和明显的滞后，以便建立一个可以推广的预测模型。

VIII. 致谢

感谢David Crandall教授提供的有益讨论和建议，感谢戴启云对本文的校对。

参考文献。

[1] H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a social network or a news media?” in *WWW*, 2010.

[2] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining,” in *LREC*, 2010.

[3] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: real-time event detection by social sensors,” in *WWW*, 2010.

[4] X. Jin, A. Gallagher, L. Cao, J. Luo, and J. Han, “The wisdom of social multimedia: Using Flickr for prediction and forecast,” in *ACM MM*, 2010.

[5] A. Tumasjan, T. Sprenger, P. Sandner, and I. Welpe, “Predicting elec-tions with twitter: What 140 characters reveal about political sentiment,” in *ICWSM*, 2010.

[6] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant, “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, 2009.

[7] H. Zhang, M. Korayem, D. Crandall, and G. LeBuhn, “Mining photo-sharing websites to study ecological phenomena,” in *WWW*, 2012.

[8] R. White, S. Dumais, and J. Teevan, “Characterizing the influence of domain expertise on web search behavior,” in *WSDM*, 2009.

[9] M. Kamvar, M. Kellar, R. Patel, and Y. Xu, “Computers and iphones and mobile phones, oh my!: a logs-based comparison of search users on different devices,” in *WWW*, 2009.

[10] G. Singh, N. Parikh, and N. Sundaresan, “User behavior in zero-recall ecommerce queries,” in *SIGIR*, 2011.

[11] ——, “Rewriting null e-commerce queries to recommend products,” in *WWW*, 2012.

[12] A. Roth and A. Ockenfels, “Last-minute bidding and the rules for ending second-price auctions: Evidence from ebay and amazon auctions on the internet,” *The American Economic Review*, 2002.

[13] H. Shah, N. Joshi, A. Sureka, and P. Wurman, “Mining ebay: Bidding strategies and shill detection,” *WEBKDD 2002-Mining Web Data for Discovering Usage Patterns and Profiles*, pp. 17–34, 2003.

[14] S. Anwar, R. McMillan, and M. Zheng, “Bidding behavior in competing auctions: Evidence from ebay,” *European Economic Review*, 2006.

[15] D. Houser and J. Wooders, “Reputation in auctions: Theory, and evidence from ebay,” *Journal of Economics & Management Strategy*, 2006.

[16] S. Guo, M. Wang, and J. Leskovec, “The role of social networks in online shopping: information passing, price of trust, and consumer choice,” in *EC*, 2011.

[17] J. Byers, M. Mitzenmacher, and G. Zervas, “Daily deals: Prediction, social diffusion, and reputational ramifications,” in *WSDM*, 2012.

[18] N. Parikh and N. Sundaresan, “Scalable and near real-time burst detection from ecommerce queries,” in *KDD*, 2008.

[19] A. Kotov, C. Zhai, and R. Sproat, “Mining named entities with tempo-rally correlated bursts from multilingual web news streams,” in *WSDM*, 2011.

[20] J. Cook, A. Das Sarma, A. Fabrikant, and A. Tomkins, “Your two weeks of fame and your grandmother’s,” in *WWW*, 2012.

[21] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch, “A word at a time: computing word relatedness using temporal semantic analysis,” in *WWW*, 2011.

[22] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos, “Identifying similarities, periodicities and bursts for online search queries,” in *SIGMOD*, 2004.

[23] A. Kulkarni, J. Teevan, K. Svore, and S. Dumais, “Understanding temporal query dynamics,” in *WSDM*, 2011.

[24] M. Mohebbi, D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi, and S. Kumar, “Google correlate whitepaper,” http://www.google.com/trends/correlate/whitepaper.pdf, 2011.

[25] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury, “How fl ickr helps us make sense of the world: context and content in community-contributed media collections,” in *ACM MM*, 2007.

[26] H. Zhang, M. Korayem, E. You, and D. Crandall, “Beyond co-occurrence: discovering and visualizing tag relationships from geo-spatial and temporal similarities,” in *WSDM*, 2012.

[27] J. Yang and J. Leskovec, “Patterns of temporal variation in online media,” in *WSDM*, 2011.

[28] J. Lin and G. Mishne, “A study of “churn” in tweets and real-time search queries,” in *ICWSM*, 2012.

[29] J. Weng and B. Lee, “Event detection in twitter,” *ICWSM*, 2011.

[30] H. Becker, M. Naaman, and L. Gravano, “Beyond trending topics: Real-world event identification on twitter,” 2011.

[31] A. Ritter, Mausam, O. Etzioni, and S. Clark, “Open domain event extraction from twitter,” in *KDD*, 2012.

[32] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, and R. Miller, “Twitinfo: aggregating and visualizing microblogs for event exploration,” in *CHI*, 2011.

[33] J. Bollen, H. Mao, and X.-J. Zeng, “Twitter mood predicts the stock market,” *Journal of Computational Science*, 2011.

[34] H. Choi and H. Varian, “Predicting the present with google trends,” *Economic Record*, 2012.

[35] S. Goel, J. Hofman, S. Lahaie, D. Pennock, and D. Watts, “Predicting consumer behavior with web search,” *PNAS*, 2010.

[36] M. Sayal, “Detecting time correlations in time-series data streams,” *Hewlett-Packard Company*, 2004.

[37] M. Kendall and A. Stuart, *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*. Hafner Publishing Company, 1961.

[38] C. W. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: Journal of the Econometric Society*, 1969.