DeepL
订阅DeepL Pro以翻译大型文件。
欲了解更多信息，请访问www.DeepL.com/pro。

预测新闻文章的受欢迎程度

Yaser Keneshloo* Shuguang Wang† Eui-Hong (Sam) Han† Naren Ramakrishnan*

摘要

消费新闻文章是我们日常生活中不可或缺的一部分，像《华盛顿邮报》（WP）这样的新闻机构在为他们的读者提供高质量的阅读体验方面花费了巨大的努力。记者和编辑面临的任务是确定哪些文章会变得流行，以便他们能够有效地分配资源，支持更好的阅读体验。新闻文章受欢迎背后的原因通常是多种多样的，可能涉及当代性、写作质量和其他潜在的因素。在本文中，我们将人气预测问题归纳为回归问题，设计了几类特征（元数据、上下文或基于内容、时间和社会），并建立了预测人气的模型。本文介绍的系统被部署在《华盛顿邮报》的一个真实环境中；我们证明了它能够准确地预测文章的受欢迎程度，其$R^2 \approx 0.8$，使用的是在发布时间的30分钟内收获的特征。

1 简介

新闻是我们日常生活中不可或缺的一部分，像《华盛顿邮报》这样的机构每天都会发布超过一千条的新闻内容。然而，并非所有这些文章都会变得同样受欢迎，因此，人气预测是记者和编辑的一个宝贵策略，可以优先考虑哪些文章需要完善。

在解决流行性预测问题中，有两个问题至关重要。首先，由于现在人们通过各种渠道消费新闻，因此有多种衡量流行度的方法，例如，WP网站上的页面浏览量，Facebook上的喜欢或分享数量，或者搜索引擎中的搜索数量。其次，文章受欢迎程度可以在本地或全球范围内定义。本地范围的衡量标准主要是为了在一个单一的新闻机构内使用，而全球范围的衡量标准则有助于确定一篇文章在其他新机构的文章中的受欢迎程度。虽然我们的最终目标是整合一系列的流行度量，包括在全球范围内，我们主要关注预测本地范围内的受欢迎程度，并使用一篇文章的页面浏览量作为其受欢迎程度的替代值。

就其性质而言，一篇新闻文章的寿命是非常短的（在其发表之后）。对《华盛顿邮报》的记者和编辑的采访表明，预测一篇文章的早期受欢迎程度比预测其长期受欢迎程度更有意义和价值。因此，在这项研究中，一篇文章的受欢迎程度被定义为文章发表后头24小时内的页面浏览量。我们将流行度预测作为一个回归问题，从一篇新闻文章中提取发表后30分钟内的特征，并使用这些特征来预测该文章在24小时内的页面浏览量。

我们的主要贡献是：

1. 我们评估了一套广泛的元数据、上下文或基于内容、时间和社会特征来预测一篇新闻文章的受欢迎程度。为了说明问题，除了文章的点击流和完整的内容之外，我们还利用了分享文章的Twitter用户的特征，估计文章新鲜度的特征，以及情感特征。

2. 我们评估了用于预测人口数量的多种回归模型，并在《华盛顿邮报》的实时系统中部署了我们表现最好的模型。

2 相关工作

新闻文章的流行度预测是一个相对新颖的问题，很少有研究解决这个问题。然而，越来越多的研究已经在预测其他类型的在线内容的受欢迎程度方面开展了。一些研究已经分析了推特上推文的传播速度或YouTube上视频的观看速度。这些研究的目标包括预测一条推文的确切转发数[26]，预测一个视频的YouTube浏览量[22，12]，估计一个Digg帖子的投票数[17]或一篇新闻文章的页面浏览数[18]，对新闻文章进行排名[24]，预测推文/新闻文章的流行范围[1]，和

*Dept. 计算机科学系，弗吉尼亚理工大学，弗吉尼亚，美国
(yaserkl,naren)@vt.edu
†ÃÃ 华盛顿 邮报》、 (shuguang.wang, sam.han)@washpost.com

Copyright © by SIAM
未经授权，禁止转载本文。
05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权；见https://epubs.siam.org/terms-privacy

预测一篇新闻文章的确切评论数[25, 24]。

对在线内容流行度的预测可以在两个阶段进行：内容发布之前或之后。有许多研究集中在使用出版后的早期测量来预测未来的成功[22,18]。另一方面，诸如[1, 25]的研究旨在预测出版前的受欢迎程度。[1]的研究使用一篇文章在Twitter上发布的次数以及一些背景特征来预测推文数量。推文中的提及次数在某种程度上可以作为流行度的替代物，但这不如文章的页面浏览量（如这里使用的）准确。在另一项研究中[27]，转发文章的用户的粉丝数量被用来预测一条推文的总转发量。26]提出了一种贝叶斯方法来预测根据两个特征，即转发者的追随者数量和转发者在转发树中的深度，来确定一条推文的转发数量。通过使用一个图形模型，该方法训练了与这些特征相关的不同参数。它使用用户的反应时间，即用户看到一个帖子和用户转发它之间的时间，作为预测最终转发数量的主要预测因素。

现有的研究将人气预测问题框定为回归[16]、分类[25, 14]，甚至是聚类[12]。这些研究中使用了各种特征来预测微博/新闻文章的受欢迎程度[23]。我们可以将这些特征分为：基于内容和时间的特征。

基于内容的特征通常是从推特/新闻文章的文本中提取的。诸如文本的情感[5, 21]、文本中的情绪[3, 2]、其语言的主观性[1]、命名实体[1]和内容的新鲜度[6, 8]等特征都被认为是与内容的病毒性高度相关的因素。[1]中的工作提出了使用发布文章的网站的类别和名称来预测病毒性的想法。图1显示了我们的WP数据集中被浏览和被推送的文章的类别分布。可以看出，在Twitter上发布的文章和在WP网站上浏览的文章，其类别和相应的分布是不同的。这表明，用户不一定与他们的朋友分享他们所读的内容。相反，他们选择特定的故事并在他们的网络中分享它们。这个结果遵循了[4]中的发现，他们发现用户只倾向于与他们的朋友分享选定的故事。

时间特征主要从以下方面提取
一篇文章的点击时间序列。在所有的时间特征中，第一小时内的转发/点击数量

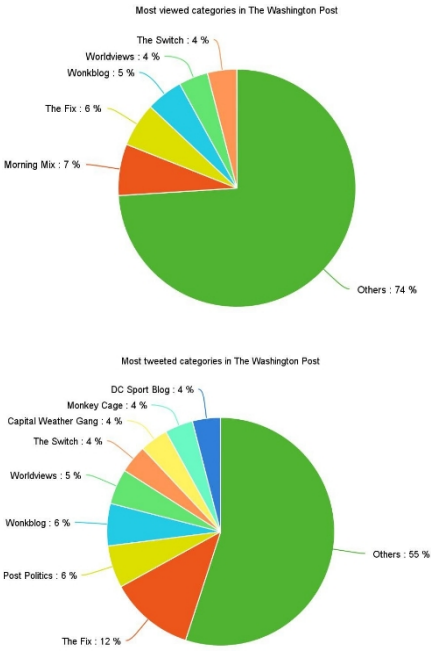

图1：《华盛顿邮报》网站上被浏览次数最多的文章（上）和被推送次数最多的文章（下）的类别分布。请注意，两者的分布有所不同。

已知发布推文/文章的时间（n_0）与t时间的最终转发/点击数高度相关，即n_t [22, 12]。根据[22]，第一小时内对数转换的转发/点击数与它的长期流行度之间存在着高度的线性相关关系。这项研究提出了一个简单的常数乘数α，根据n_0来估计n_t，即$\log n_t = \alpha \cdot \log n_0$。（我们使用类似的方法作为我们的基线模型。）[20]提出了这种方法的扩展，他们用定期（15分钟）到第一小时的样本来代替n_0。然而，在这种方法中，一篇文章的点击量在所有同时发布的其他文章中的相对重要性被忽略了。(我们通过建议使用规范化的页面浏览功能来改进这一功能，该功能是相对于在同一时间发布的所有文章而言的。)除了这些特征外，转发加速和转发树中的转发深度也被用于[14]预测Twitter消息的受欢迎程度。

3 人气预测

在这一节中，我们介绍了我们提出的预测一篇新闻文章受欢迎程度的方法。我们的目标是预测一篇新闻文章的页面浏览量。

Copyright © by SIAM
未经授权，禁止转载本文。

05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权；见https://epubs.siam.org/terms-privacy

在文章发表后的第一天就会收到。我们在所有文章发表后跟踪30分钟，并提取一系列的时间、社会和背景特征进行预测。

3.1 元数据功能。WP元数据包含每篇文章的详细信息，如标题、完整内容、关键词、作者、类型（博客或文章）、类别、新闻部分和出版日期。从一篇文章的发表日期中，我们提取了文章发表的时间和星期。除了这些特征外，我们还使用作者姓名、新闻类型、类别和栏目作为我们的元数据特征。

3.2 基于内容的特征。情景特征主要涉及文章的标题、关键词和内容。我们把这些上下文特征分成两组：第一组包括从WP元数据数据集中提取的特征，第二组是通过查询我们的WP数据集的*伪*档案而提取的（后面会详细解释）。

3.2.1 情感。作为另一个上下文特征，我们将文本片段的情感提取为属于正面、负面、中立或Com-pound类别的概率。我们使用Vader情感分析器[13]来实现这一目的。除了文本的情感之外，文本的情感也是影响病毒性的一个重要因素[3]。然而，正如[3]中提到的，没有语言学工具可以捕捉文本中的情感，我们必须利用人力来确定文本中的情感。我们创建了两个不同的情感特征集，一个是完整内容的情感，另一个是新闻文章的标题。我们使用这两个不同的集合是因为目前大多数的情感分析器在对长文本进行分类时都面临着一个问题。

3.2.2 命名实体的提取。命名实体也是影响新闻文章病毒性的一个重要因素。关于一个知名的地方或一个人的新闻文章可以为自己带来大量的关注。我们为此使用了斯坦福NLP库[1]。虽然只捕获命名实体的数量可能无法捕捉到每个实体的重要性，但它将提供一个指标，说明文章对一个主题的谈论有多彻底。在我们的实验中，我们将展示这些数字如何帮助改善我们的预测。
(正如后面所描述的，我们也捕捉到了以下的重要性
命名的实体，通过估计*伪的*计数来实现。
数据集）。

[1]http://nlp.stanford.edu/software/

3.2.3 可读性。Berger和Milkman[3]声称，较长的文章往往会被更多地分享。相反，我们发现，一篇文章的长度和它所收到的页面浏览量之间的相关性很小。3]中的说法是基于这样一个前提：当记者在写热点话题时，他们倾向于写更长的文章。与文章长度一起，文章的可读性也是一个因素，为此有几个不同的衡量标准（通常基于估计一个人理解文本所需的教育年限）。这些方法大多使用字和句子的长度、复杂词汇的数量以及文本内的音节数的组合来估计可读性。为了我们的目的，我们使用Flesch-Kincaid可读性测试、Gunning-Fog得分、自动可读性指数和Coleman-Liau指数来确定可读性（关于这些衡量标准的描述见[10]）。除了这些指标外，我们还考虑了句子的数量、复杂词的数量、复杂词占总词数的百分比、音节的数量、每个词的平均音节数、每句的平均词数、平均句子长度、以及文章的标题和全文内容的长度。

3.2.4 新鲜度。所有上述的特征都是用标准的方法和软件来测量的。在这一节中，我们提出一种方法，旨在确定一篇文章的新鲜度。如前所述，病毒性新闻文章通常由新鲜和令人惊讶的信息驱动。为了捕捉一篇文章的新鲜度，我们必须对近期来自其他新闻机构的文章进行建模。获取这些信息超出了这项工作的范围；我们创建了一个*伪*数据集，包含过去（在本文中，2014年9月之前）发表的、至少获得一次页面浏览的WP文章（在2014年9月至2015年3月期间）。伪档案使用Lucene对标题、完整内容和关键词进行索引，我们使用Lucene内置的评分机制来确定前10名最相似的文章。从相似性结果中，我们提取了以下特征：

- **主题交集。**被查询的文章和存档的文章之间的主题交集被定义为：：

$$TI = \frac{|keywords_q \cap keywords_a|}{|keywords_q \cup keywords_a|} \quad (3.1)$$

其中$keywords_q$和$keywords_a$分别是被查询文章和前十名文章中的关键词集合。这个特征将接近于

Copyright © by SIAM
未经授权，禁止转载本文。

05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权；见https://epubs.siam.org/terms-privacy

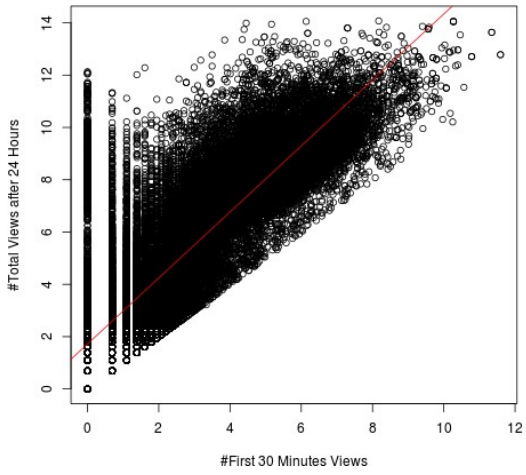

图2：对数转换后的前30分钟页面浏览量与24小时后WP文章的总页面浏览量之间的相关性。红线代表前30分钟页面浏览量的简单线性回归估计。这个估计值周围的明显散点表明，为了提高性能，需要额外的功能。

新的或新的文章为零，旧的故事接近1。

- **前十个故事的页面浏览量。**对于前十篇类似的文章中的每一篇，我们找出该文章在被查询文章的出版日期之前所获得的页面浏览量。我们将这些特征视为TC_i，对$i = 1, - - -, 10$进行排序，使$TC_i < TC_{i+1}$。这些特征将为我们提供对新文章的病毒性的粗略估计
与早期类似物品的性能相比。

- **前十名的故事内容相似度。**对于每个
在前十篇文章中，我们发现余弦模拟
它的内容与被查询文章的内容的相似性。这个特征也反映了内容的新鲜程度。我们将这些特征表示为TS_i，$i = 1, - - -, 10$排序，使$TS_i < TS_{i+1}$。

除了上述特征外，我们还捕捉到与被查询文章相似的文章总数。我们把这个特征称为*命中率*，它代表了被查询文章在我们存档数据库中的命中率。这个数值越低，文章的内容就越新鲜。

3.3 时间特征。我们以5分钟的间隔从页面浏览的点击流数据中提取时间特征。在分析YouTube的浏览量时，已经发现第一小时的对数转换的浏览量和一个条目的最终浏览量之间存在高度的相关性[22]。在这项工作中，前30分钟的页面浏览量被认为是预测的主要特征。(其他工作，如[14]，使用额外的特征，如前30分钟的转发时间序列和转发加速。）在对新闻文章进行类似的分析时，如图2所示，我们发现虽然类似的关系近似成立，但观察到的数据在空间上更加分散，这表明这种估计最多只能作为一种约束。在这里，我们提出了额外的特征来预测页面浏览量：

- **发布时间和第一页浏览之间的时间差。**这个特征捕捉了人们对一篇新闻文章的反应速度。我们也称这个特征为*页面浏览反应时间*。我们预计，对于病毒性文章，这个数字会很小，而对于普通文章，这个数字会更大。在我们的WP数据集中，所有文章的中位时间差是237分钟。85%的前1%的病毒性文章的反应时间少于200分钟。

- **30分钟后的页面浏览数。**这个数字反映了一篇文章在发表后的前30分钟内收到的页面浏览总数。平均而言，前1%的文章在发布后收到约571页的浏览量。
30分钟，一天后为30万。值得一提的是，其余文章前30分钟的平均页面浏览量为18左右。

- **页面浏览加速。**我们使用来自[14]的方法来捕获页面视图加速：

$$\Sigma_N \quad x \quad x$$

$$\text{(3.2)} \qquad \text{加速} = \frac{t=2\ nt - \text{nt}^{-1}}{N}$$

其中n^x_t是文章x在时间间隔t的页面浏览量，N是前30分钟内的时间间隔总数。（由于我们使用5分钟的时间间隔来建立我们的时间序列，所以前30分钟的$N = 6$）。

- **页面浏览时间序列。**与[14]类似，我们使用页面浏览时间序列的每个时间区间的值，即n_t，作为预测因素。

- **归一化的页面浏览时间序列。**我们将页面浏览时间序列与同一时间发表的其他文章的时间序列进行了归一化。

Copyright © by SIAM
未经授权，禁止转载本文。

05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权：见https://epubs.siam.org/terms-privacy

因此，给定在同一时间发表的m篇文章，归一化的时间序列如下：

$$NC_t = \frac{n_t^x}{\sum_{i=1}^{m} n_t^i} \quad (3.3)$$

其中n_t^x是文章x在时间间隔t内，n_t^i是总的i^{th}文章的页面浏览量，该文章与x文章同时发布。我们也可以使用一个时间窗口来对这些时间序列进行归一化，即对过去几小时内发表的所有文章进行归一化。 归一化计数是一个比计数本身更好的衡量标准，因为它可以发现一篇文章在同一时间发表的其他文章中的相对重要性。在给定*的*几篇文章中，得到更多关注的文章会有一个更高的归一化计数。因此，一个接近1的归一化计数意味着该文章在同一时间发表的所有其他文章中受到更多的关注。

3.4 社会媒体特征。利用点击流数据集，我们生成另一个数据集来捕捉与每篇文章相关的社会媒体（Twitter）活动。我们使用Topsy API[2] 来提取所有分享WP网址的推特。对于每篇文章，我们创建一个类似于点击流数据集的推文和转发时间序列，也就是说，这个数据集的每一行都包含推文的时间戳和它的推文ID。利用每条推文的TweetID，我们使用其API[3] 查询Twitter，以访问用户资料并生成一个用户资料数据集。对于每个用户，我们提取关注者的数量，列出的数量，朋友的数量，以及状态信息/更新的数量。对于每篇新闻文章，从提取的与每篇文章相关的推文中，我们建立转发时间序列。我们使用一篇文章的转推时间序列来生成另一组社会媒体驱动的时间特征：

- **发布时间和第一条推文之间的时间差。**我们把这个特征称为*推文反应时间*，和*页面浏览反应时间*一样，推文反应时间可以捕捉到人们在Twitter上分享新闻的速度。与*页面浏览*的*反应时间*类似，我们预计病毒性文章的分享速度会比其他新闻文章快。在我们的WP数据集中，对于在Twitter上至少有一条推文的文章，推文反应时间的平均值和中位数是13分钟。前1%的病毒性文章中，83%的文章的推文反应时间少于200分钟。

[2]http://api.topsy.com/[3] https://dev.twitter.com/

- **30分钟后的转发数量。**这个数字是指一篇文章在发表后30分钟内收到的转发总数。

  平均而言，前1%的文章收到了约106
  30分钟后转发。令人惊讶的是，其余的
  在这之后，文章平均收到约194次转发。
  30分钟。这种行为的原因是：、

  正如[15]中提到的，人们使用社交媒体来分享他们周围发生的几乎所有事情。然而，这些新闻文章中只有一小部分会成为病毒式传播。我们以后会在第4节中看到，与前30分钟的页面浏览量不同，转发数对预测性能没有明显的贡献。

- **30分钟追随者"的数量。**我们提取所有分享文章的用户的追随者数量，并将其汇总，找到这个数字。这个特征可以捕捉到在前30分钟内接触到该文章的大致人数。平均而言，分享前1%文章的用户的粉丝总数约为380人，而对于其他文章，这一数字为14。这表明，平均而言，分享病毒性文章的人往往有更多的追随者。

- **转发/关注者加速**：与页面浏览加速类似，该功能使用类似于方程3.2的公式计算。

- **转发/关注者时间序列**：我们使用转发/关注者时间序列的每个时间区间的值，即n_t，作为预测因素。

- **归一化的转发/关注者时间序列**：这个功能的计算方法也类似于我们计算规范化页面浏览时间序列的方法。

表1总结了我们研究中使用的所有特征。

4 实验

总的来说，我们为每篇文章提取了105个特征。这些特征被组织成几组，并评估了它们比前面描述的基线方法的增量。所有的实验都是通过10倍的交叉验证进行的，我们使用平均调整后的R^2（$AdjR^2$）值来报告性能。除了整体性能，我们还对模型在病毒性文章上的性能感兴趣。因此，在运行的每个折叠中，我们首先在训练集上训练模型，然后在两个测试集上测试：一个包含该折叠中的所有测试文章，另一个只包含第一个测试集中最受欢迎的文章的1%。

Copyright © by SIAM
未经授权，禁止转载本文。

05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权：见https://epubs.siam.org/terms-privacy

表1：本工作中评估的特征。

元数据功能	
文章类型	无论该文章是博文还是文章。
文章类别	不同的人看待不同的类别； 一些类别通常不会产生病毒性文章； 其他人则产生更多的热门新闻。
第1条	与文章的类别类似，文章的章节也是影响其受欢迎程度的一个重要因素。 大多数发表的文章属于体育、政治和意见部分。
出版日期	我们记录这一特征是在一天中的时间和 文章发表的星期的日期。
作者姓名	我们的数据集中有超过12000名作者；瓦莱丽-施特劳斯和丹-斯坦伯格等作者。 发表文章最多。
背景特征	
感受	我们为标题和主要文章提取情感分数。情感被定义为四个类别的概率： {负数、正数、复数、中性}。
命名的实体	文章中的人物、地点和组织的数量
文本的可读性	捕捉到文件可读性的不同方面的五种措施
文章的新鲜度	整理成： (i) 主题交叉； (ii) 10篇最相似文章的点击数； (iii) 10篇最相似文章的内容相似度；以及 (iv) 历史数据集中类似文章的数量
时间特征	
首次查看时间差	发布时间和第一页浏览之间的时间差。
第一个30分钟视图	发布30分钟后的页面浏览量。
网页浏览加速	一篇文章在前30分钟内被阅读的比率。
页面查看时间序列 归一化的页面浏览时间序列	前30分钟的时间序列，以5分钟为间隔组织的意见。
社会功能	
第一条推文时间差	发布时间的时间差 和第一条推特。
第一个30分钟的推文量	发布30分钟后的推文数量。
第一个30分钟的追随者人数	在30分钟内发布新闻的用户的追随者数量。
鸣叫/追随者加速	一篇文章在前30分钟内被推送的速度。
推特/关注者时间序列 归一化的推特/关注者时间序列	前30分钟的推文的时间序列，以5分钟为间隔组织。

Copyright © by SIAM
未经授权，禁止转载本文。

05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权；见https://epubs.siam.org/terms-privacy

表2：在完整（测试）数据集和前1%（病毒）数据集上不同回归模型的比较。

模型	完整($Adj R^2$)	前1% (R^2)
多线性回归	79.4	78.2
LASSO回归	72	52.1
山脊回归	80.3	54.5
树状回归	82.9	42.5

4.1 模型选择。表2比较了多个回归模型在我们数据集上的表现。如图所示，基于树的回归在完整数据集上表现最好，但在虚拟数据集上表现很差。另一方面，多元线性回归（MLR）模型[11]在两个测试数据集上都有令人满意的表现，我们在其余的实验中使用这种方法。

基线方法利用最强的信号来预测页面浏览量，即文章发表后前30分钟内的对数转换后的页面浏览数。我们的数据集包含了2014年9月至10月间发表的超过41K篇文章，其中37K篇文章的元数据信息是可用的。因此，在我们的数据集中，大约有4K篇文章的元数据特征缺失，我们用0作为默认值。此外，对于那些没有被推送的文章，我们使用自定义的默认值来填补缺失的特征。例如，对于文章的第一条推文和发表日期之间的时间差，我们对没有这个测量值的文章使用一个非常大的值。对于其他社会特征，如30分钟内的转发/关注者数量和转发/关注者时间序列，我们使用零。下面的结果表明，我们的模型在处理缺失值方面有足够的稳定性。

表3显示了我们回归分析的结果
在这个数据集上。该表中的粗体条目显示了对基线模型提供最佳提升的特征集。在其他特征集中，只添加元数据特征提供了最大的提升。如果我们使用完整的特征集，我们将基线方法的性能提高10%。为了说明使用时间特征所取得的小幅提升也是一种显著的提升，我们使用配对t检验来检验结果的显著性，发现使用时间特征所取得的提升是极其显著的，p值<0.0001。此外，为了更好地了解新鲜度特征对模型性能的影响，我们将这些特征从内容特征中删除。 虽然

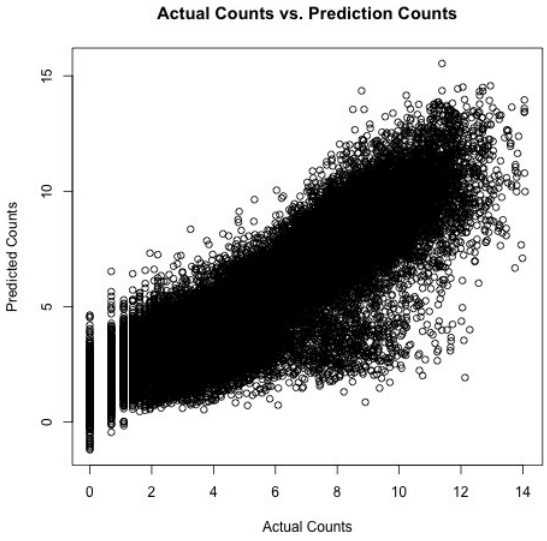

图3：实际页面浏览量与预测值的关系图。

表3：完整数据集和前1%数据集上的回归模型的调整后的R^2 和R^2 值。

模型	完整($Adj R^2$)	前1% (R^2)
基准线	69.4	74
基线+时间性	70.4	72.1
基线+社会	72.5	77.3
基线+内容	71.1	**79.3**
基线 + 内容 - 新鲜度	70.6	79
基线+元数据	**77.2**	78.1
所有功能	79.4	78.2

在没有新鲜度特征的情况下，只导致了性能的轻微下降，根据t检验，这种差异是极其显著的。

为了了解我们的模型在最具病毒性的文章上的表现，我们用我们训练好的模型来预测这组文章的页面浏览量。在这个实验中，由于特征的数量超过了样本的数量，我们使用R^2 的措施来报告性能。表3显示了我们的模型在这个数据集上的表现。这个数据集的基线R^2 是69.4，而使用完整的特征集比基线有4%的提升。根据表3，尽管元数据信息对$AdjR^2$ 分数的提升最大，但基于内容的特征对预测病毒性文章的帮助最大。图3显示了24小时后的实际点击数与我们整个模型在完整数据集中的预测值的对比图（与图2比较）。

Copyright © by SIAM
未经授权，禁止转载本文。
05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权；见https://epubs.siam.org/terms-privacy

4.2 重要特征。为了量化我们数据集中的重要特征，我们分别探索每一组特征，以确定在模型性能中提供最大提升的特征集。我们的目标是提取能提供最大$AdjR^2$值的最佳特征子集。表4和表5分别描述了对完整数据集和前1%数据集的实验结果。根据这个实验，在时间特征中，页面浏览时间序列和其归一化时间序列是两个数据集中最重要的特征。请注意，当我们只使用这两个特征来预测前1%数据集中的页面浏览量时，我们得到的R^2值比使用全套时间特征时要高。这是由于时间特征集中的一些特征之间可能存在多重共线性。

同样，重要的社会特征也是相同的
完整的数据集和前1%的数据集之间。在所有的社会特征中，第一条推文和文章发表日期之间的时间差、文章发表后30分钟内分享文章的用户的粉丝数量、转发时间序列和其归一化时间序列是最重要的特征。可以看出，使用这些选定的特征的性能与表3的结果完全相同。
尽管如表4和表5所示，完整的数据集和前1%的数据集有相似的社会和时间特征，但它们有一组不同的重要背景特征。根据我们的实验，在完整的数据集中，前十个故事的页面浏览量和con- tent相似度的子集是由子集选择过滤器挑选的。此外，在所有的情感特征中，只有内容中立性和复合性的概率在预测中被认为是重要的。另外，从四个不同的可读性衡量标准中，SMOGIndex被选为最佳子集选择方法。另外，尽管已知的说法是文章长度是预测流行度的一个好属性[3]，但根据我们的分析，标题长度是有助于提高性能的特征。
然而，在前1%的数据集中，主题交集
是重要的特征之一。此外，内容的可读性不是一个重要的特征，没有一个可读性特征被用于这个数据集。从表5可以看出，在所有的文本特征中，所有提出的与文章新鲜度有关的特征都被最佳子集选择方法选中。这放大了文章新鲜度在预测文章病毒性方面的重要性。

4.3 在《华盛顿邮报》的部署 在本文中，我们探索了各种特征并训练了一个

表4：$AdjR^2$值和使用最佳子集选择方法从每个特征集中提取的重要特征列表（所有文章）。

	锛滴層閔庶黩鑰咂悕锛滴層閔庶黩鑰咂悕	**重要特点**
时间性	70.3	页面浏览量和归一化页面浏览量的时间序列
社会	72.5	推特时差，#关注者在前30分钟、 转发和归一化转发的时间序列
内容	70.9	前十名故事的页面视图，前十名故事的内容相似度、 标题长度，内容中立的概率 和复合度，以及SMOGIndex

表5：R^2值和使用最佳子集选择方法从每个特征集中提取的重要特征列表（病毒性文章）。

	R2	**重要特点**
时间性	72.6	页面浏览量和归一化页面浏览量的时间序列
社会	77.3	推特时差，#关注者在前30分钟、 转发和归一化转发的时间序列
内容	79.4	主题交集，前十名故事页面浏览，前十名故 事内容相似度，标题长度、 和内容中立性

在流行度预测任务中使用了回归模型。为了帮助《华盛顿邮报》的记者和编辑，我们部署了这个模型并为每篇文章建立了一个实时的预测系统。一旦一篇新闻文章发表，预测系统就开始跟踪它并提取本文所述的特征。在Splunk[4]的帮助下，我们建立了一个仪表板，根据预测的受欢迎程度来排列文章。当我们对新闻文章进行预测时，我们也会跟踪文章的实际页面浏览量以进行评估。除了为编辑和记者提供人气预测外，我们还能确定我们提出的回归模型在最新的新闻文章中的表现如何。我们使用同样的设置来评估预测模型的性能。我们收集所有在2015年8月（预测系统部署后）发表的文章。表6总结了这个模型的性能。虽然部署的模型没有使用所有提议的功能，但它的性能与表3中显示的结果相当。

5 总结

本文是第一个预测新闻文章页面浏览量的努力。我们探索了对新闻文章受欢迎程度起重要作用的不同因素，即时间性、社会性和背景性特征。我们的评估结果表明，在这三组特征中，与文章的新鲜度有关的背景特征是预测病毒性文章页面浏览量的最重要因素、

[4]http://www.splunk.com/

Copyright © by SIAM
未经授权，禁止转载本文。

05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权；见https://epubs.siam.org/terms-privacy

表6：对2015年8月发表的文章的部署模式的评价

数据	*锛屾垚鍥而不舍*[2]	前1% $(R)^2$
2015年8月	0.829	0.620

而元数据特征则是预测新闻文章总体表现的最强信号。从这三组特征中，我们还确定了对新闻文章的受欢迎程度预测有显著贡献的最有效的特征。在优秀的离线评估结果的激励下，我们在《华盛顿邮报》部署了该模型。未来工作的目标不仅仅是人气预测，还包括支持文章在生命周期中的创建、发表和修改等其他方面。

参考文献

[1] Bandari, R.; Asur, S.; and Huberman, B. A. *The pulse of news in social media：预测人气。CoRR* abs/1202.0332, 2012.

[2] Berger, J.和Milkman, K.*社会传播、情绪和感。争论，以及在线内容的病毒性。沃顿大学再搜索论文*，2010年。

[3] Berger, J., and Milkman, K. L. *What makes online con. tent viral?营销研究杂志*，49（2）：192-205，2012。

[4] Berger, J., and Schwartz, E. M. *What drives immediate 和持续的口碑？营销研究杂志》*，48(5):869-880，2011。

[5] Berger, J. *唤醒增加了社会传播的…。形成。心理科学杂志*，22（7）：891- 893，2011。

[6] Borghol, Y.; Ardon, S.; Carlsson, N.; Eager, D.; and Mahanti, A. *The untold story of the clones: content-agnostic factors that impact youtube video popularity.* In *Proceedings of the SIGKDD'12*, 1186-1194, 2012.

[7] Castillo, C.; El-Haddad, M.; Pfeffer, J.; and Stempeck、M.*使用社交媒体的反应来描述在线新闻故事的生命周期。*In *Proceedings of the CSCW'14*, 211-223, 2014.

[8] Cha, M.; Kwak, H.; Rodriguez, P.; Ahn, Y.-Y.; and Moon, S. *Analyzing the video popularity char. 大规模用户生成内容系统的特性 tems.IEEE/ACM网络交易（TON）。* 17(5):1357-1370, 2009.

[9] Cherkasova, L., and Gupta, M. *Analysis of enterprise media server workloads: access patterns, locality, con. 帐篷的演变，以及变化的速度。IEEE/ACM Trans-actions on Networking* 12(5):781-794, 2004.

[10] DuBay, W. H. *The Principles of Readability.*在线提交（2004年）。

[11] Freedman, D. *Statistical Models：Theory and Practice*. 剑桥大学出版社，2005年。

[12] Gill G.; Crovella, M.; and Matta, I. *描述和预测视频访问模式。在IEEE INFOCOM'11会议上*，16-20，2011。

[13] Hutto, C. J., and Gilbert, E. *VADER: A parsimonious 基于规则的社交媒体文本情感分析模型。*In *Proceedings of the ICWSM'14*, 2014.

[14] 孔繁森, Y. F., 和冯, L. *预测未来. 微博中的转发计数。Journal of Computa- tional Information Systems* 10(4):1393-1404, 2014.

[15] Kwak, H.; Lee, C.; Park, H.; and Moon, S. *What Twitter是一个社交网络还是一个新闻媒体？* 在 *WWW'10会议论文集*，591-600，2010。

[16] Lee, J. G.; Moon, S.; and Salamatian, K. *Modeling and predicting the popularity of online contents with cox比例危险回归模型.Journal of Neurocomputing* 76(1):134-145, 2012.

[17] Lerman, K., and Hogg, T. *Using a model of social 动态预测新闻的受欢迎程度。*In *Proceedings of WWW'10*, 621-630, 2010.

[18] Marujo, L.; Bugalho, M.; Neto, J. P. d. S.; Gershman, A.; and Carbonell, J. *Hourly traffic prediction of news stories. arXiv preprint arXiv: 1306.4608*, 2013.

[19] Mishne, G., and De Rijke, M. *A study of blog search.*In *Advances in information retrieval*.Springer.289-301, 2006.

[20] Pinto, H.; Almeida, J. M.; and Gonçalves, M. A. *Using 预测Youtube视频的流行程度的早期浏览模式。*In *Proceedings of the WSDM'13*, 365-374, 2013.

[21] Reis, J., Benevenuto, F., Olmo, P., Prates, R.、Kwak, H., and An, J. *Breaking the News：arXiv preprint arXiv:1503.07921*, 2015.

[22] Szabo, G., and Huberman, B. A. *Predicting the popu-A. 在线内容的稀有性。ACM的通讯* 53(8):80-88, 2010.

[23] Tatar, A.; De Amorim, M. D.; Fdida, S.; and Antoniadis, P. *A survey on predicting the popularity of web content.互联网服务与应用杂志》*5（1）：1-20，2014。

[24] Tatar, A., Antoniadis, P., De Amorim, M. D., and Fdida, S. *Ranking news articles based on popularity prediction.*In *Proceedings of ASONAM'12*, 106-110, 2012

[25] Tsagkias, M.; Weerkamp, W.; and De Rijke, M. *预测在线新闻故事的评论量。*In *Proceedings of the CIKM'09*, 1765-1768, 2009.

[26] Zaman, T.; Fox, E. B.; Bradlow, E. T.; et al. *A bayesian 预测推文流行度的方法.The Annals of Applied Statistics* 8(3):1583-1611, 2014.

[27] Zhao, Q.; Erdogdu, M. A.; He, H. Y.; Rajaraman、A.; and Leskovec, J. *SEISMIC: A self-exciting point process model for predicting tweet popularity.CoRR* abs/1506.02594, 2015.

Copyright © by SIAM
未经授权，禁止转载本文。

05/22/23下载到202.119.41.235.再分发须经SIAM许可或版权；见https://epubs.siam.org/terms-privacy