使用搜索引擎查询数据检测流感流行病

杰里米金斯伯格[1]，马修H。莫赫比[1]，拉詹S。帕特尔[1]林妮特·布拉默[2]马克艾伦斯莫林斯基[1]& Larry辉煌[1]

[1]谷歌公司。[2]美国疾病控制和预防中心

季节性流感的流行是一个主要的公共卫生问题，每年在全世界造成数千万例呼吸系统疾病和25万至50万人死亡[1]. 除季节性流感外，一种以前没有免疫力的新型流感病毒毒株表明人传人可能导致数百万人死亡的大流行[2]. 及早发现疾病活动，然后迅速采取反应，可以减少季节性流感和大流行性流感的影响[3,4]. 改进早期检测的一种方法是以在线网络搜索查询的形式监测寻求健康的行为，每天由全世界的数百万用户提交这些查询。在这里，我们提出了一种分析大量谷歌搜索查询的方法来跟踪人群中的流感样疾病。因为某些查询的相对频率与医生访问高度相关的百分比的病人出现流感样症状，我们可以准确地估计当前水平的每周流感活动在美国的每个地区，报告大约一天的滞后。这种方法可以使利用搜索查询来检测具有大量网络搜索用户的地区的流感流行病成为可能。

本文最初发表于2009年2月19日《自然》第457卷，
doi:10.1038/nature07634

http://dx.doi.org/10.1038/nature07634

传统的监控系统，包括美国使用的那些系统。S.美国疾病控制和预防中心（CDC）和欧洲流感监测计划（EISS）都依赖于病毒学和临床数据，包括流感样疾病（ILI）的医生就诊。CDC每周发布来自这些监测系统的国家和地区数据，通常报告滞后1-2周。

为了提供更快的检测，已经建立了创新的监测系统，以监测流感活动的间接信号，如呼叫量到电话分类咨询线[5]以及非处方药的销售[6].据信，每年约有9000万美国成年人在网上搜索有关特定疾病或医疗问题的信息[7]，使得网络搜索查询成为关于健康趋势的独特有价值的信息来源。此前利用在线活动监测流感的尝试已经计算了提交给瑞典医疗网站的搜索查询[8]，访问者在美国的某些页面。健康网站[9]，用户点击加拿大的搜索关键字广告[10].一组包含“流感”或“流感”一词的雅虎搜索查询被发现与多年来的病毒学和死亡率监测数据相关[11].

我们提出的系统建立在这些早期工作的基础上，利用一种发现流感相关搜索查询的自动方法。通过处理来自5年谷歌网络搜索日志的数百亿次个人搜索，我们的系统生成了更全面的模型，用于流感监测，并对美国流感样疾病的区域和州一级（ILI）活动进行了估计。在线搜索引擎在全球的广泛使用可能使模型最终在国际环境中开发。

通过汇总2003年至2008年之间提交的在线网络搜索查询的历史日志，我们计算了美国5000万个最常见的搜索查询的每周计数的时间序列。为每个状态中的每个查询保留了单独的聚合每周计数。没有保留任何关于任何用户的身份的信息。每个时间序列都被标准化，通过将特定一周内的每个查询的计数除以一周内在该位置提交的在线搜索查询的总数，从而得到一个查询分数（补充图1）。

我们试图开发一个简单的模型，估计在特定地区随机的医生就诊与流感样疾病（ILI）相关的概率；这相当于与ILI相关的医生就诊的百分比。使用了一个单一的解释变量：从同一区域提交的随机搜索查询与ili相关的概率，由下面描述的自动方法确定。我们使用一个ILI医生就诊的对数概率和一个与ILI相关的搜索查询的对数概率来拟合一个线性模型：

$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \varepsilon$$

其中，P为ILI医生就诊的百分比，Q为与ILI相关的查询分数，β_0是拦截，β_1为乘法系数，ε为误差项。logit (P)是P/（1-P）的自然对数。

来自美国疾控中心的公开历史数据。流感哨点提供者监测网络[12]被用来帮助建立我们的模型。在美国的9个监测区域中的每一个，美国疾病控制与预防中心报告了每周与ili相关的哨兵提供人员的所有门诊就诊的平均百分比。在年度流感季节以外的几周内没有提供任何数据，我们从模型拟合中排除了这些日期，尽管我们的模型被用于生成这几周的未经验证的ILI估计数。

我们设计了一种自动选择ili相关搜索查询的方法，不需要关于流感的先验知识。我们测量了如果我们只使用一个查询作为解释变量Q，我们的模型在每个区域的CDC ILI数据的有效性。我们以这种方式对我们数据库中的5000万个候选查询分别进行了测试，以确定能够最准确地模拟每个地区的CDC ILI访问百分比的搜索查询。我们的方法奖励查询显示区域差异类似于CDC ILI数据：随机搜索查询可以适合ILI百分比在所有九个地区远远低于随机搜索查询可以适合一个位置（补充图2）。

自动查询选择过程产生了一个得分最高的搜索查询列表，根据横跨9个区域的平均z转换相关性进行排序。为了决定哪些查询将包含在与ili相关的查询分数Q中，我们考虑了N个评分最高的不同查询集。我们根据每个集合中查询的总和来测量这些模型的性能，并选择了N，这样我们就获得了对9个区域的样本外ILI数据的最佳拟合（图1）。

结合N个=45个得分最高的查询，可以获得最佳拟合。这45个搜索查询，虽然已经被选中了

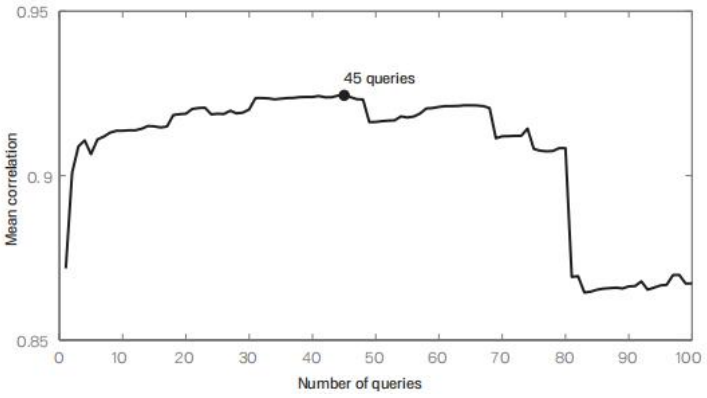

图1：评估在与ili相关的查询部分中包含多少个得分最高的查询。通过将前45个搜索查询相加，获得了在交叉验证过程中估计样本外点的最大性能。在添加查询81后，即“奥斯卡提名”。

自动地，似乎始终与流感类似的疾病相关。前100名中的其他搜索查询，不包括在我们的模型中，包括“高中篮球”等主题，这些主题往往符合美国的流感季节（表1）。

使用这个与ILI相关的查询分数作为解释变量，我们将一个最终的线性模型拟合到2003年至2007年期间所有9个区域的每周ILI百分比，从而学习一个单一的、区域独立的系数。该模型能够与cdc报告的ILI百分比的良好拟合，平均相关性为0.90（min=0.80，max=0。96，n个=9个区域）（图2）。

最终的模型在2007-2008年之前未测试的数据的每个区域的42个点上进行了验证，这些数据被排除在之前的所有步骤之外。对这42个点产生的估计数得到的平均相关性为0.97（min=0.92，max=0。99，n=9区域)与cdc观察到的ILI百分比。

在整个2007-2008年流感季节，我们使用我们的模型的初步版本来生成ILI估计数，并每周与CDC流感部门的流行病学和预防部门分享我们的结果，以评估及时性和准确性。图3说明了整个季节中不同时间点的可用数据。在这9个地区，我们能够在美国CDC中心发布报告前1-2周持续估计当前的ILI百分比。流感哨点提供者监测
网络

由于局部流感监测对公共卫生规划特别有用，我们试图进一步验证我们的模型，与个别州的每周ILI百分比相比。CDC没有公开州一级的数据，但我们根据犹他州提供的州报告的ILI百分比验证了我们的模型，并在42个验证点上获得了0.90的相关性（补充图3）。

谷歌网络搜索查询可用于准确估计美国九个公共卫生区域的流感样疾病百分比。由于搜索查询可以快速处理，因此得到的ILI估计值始终比CDC的ILI监测报告早1-2周。这种方法提供的早期发现可能成为抵御美国未来流感流行的一条重要防线，并可能最终在国际环境中实现。

最新的流感估计数可以使公共卫生官员和卫生专业人员能够更好地应对季节性流行病。如果一个地区的ILI医生就诊人数早期急剧增加，就有可能将额外的资源集中在该地区，以确定疫情的病因，提供额外的疫苗能力或在必要时提高当地媒体的认识。

前45个查询，接下来的55个查询

搜索查询主题	N	加权	N	加权
流感并发症	11	18.15	5	3.40
感冒/流感治疗方法	8	5.05	6	5.03
一般流感症状	5	2.60	1	0.07
流感的术语	4	3.74	6	0.30
特异性流感症状	4	2.54	6	3.74
流感并发症的症状	4	2.21	2	0.92
抗生素药物治疗方法：	3	6.23	3	3.17
一般流感疗法	2	0.18	1	0.32
一种相关疾病的症状	2	1.66	2	0.77
抗病毒药物治疗	1	0.39	1	0.74
相关疾病	1	6.66	3	3.77
与流感	0	0.00	19	28.37
无关				
	45	49.40		55
		50.60		

表1：在搜索查询中发现的与CDC ILI数据最相关的主题。在我们最终的模型中使用了前45个查询；下面的55个查询将被用来进行比较。表示每个主题中的查询数量，以及查询体积加权计数，这反映了每个主题中查询的相对频率。

该系统并不是为了取代传统的监测网络，也不是为了取代基于实验室的诊断和监测的需要。显著增加
与ili相关的搜索活动可能表明需要进行公共卫生调查，以确定所涉及的病原体或多个病原体。人口数据通常由传统监测提供，不能通过搜索查询获得。

如果出现引起大流行的流感毒株，准确和早期发现ILI百分比可能使公共卫生官员能够采取更有效的早期应对措施。虽然我们不能确定搜索引擎用户在这种情况下会如何行为，但受影响的个人可能会提交在我们的模型中使用的相同的与ili相关的搜索查询。另外，健康个体中的恐慌和担忧可能会导致与ILI相关的查询分数的激增，并导致对正在进行的ILI百分比的夸大估计。

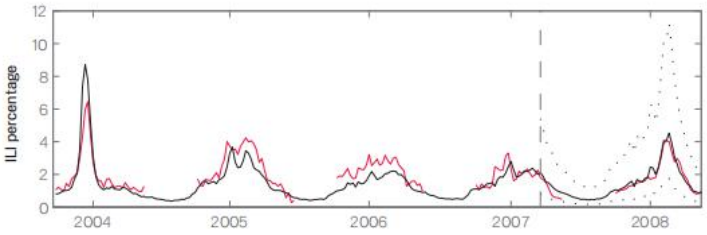

图2：中大西洋地区的模型估计值（黑色）与cdc报告的ILI百分比（红色）的比较，包括模型拟合和验证的点。从模型拟合的该区域获得超过128个点的相关性为0，而相关性为0。.85通过超过42个验证点，获得了96个验证点。表示有95%的预测区间。

当然，我们的模型中的搜索查询并不是专门由经历流感样症状的用户来提交的，我们观察到的相关性只在大人群中有意义。尽管存在很强的历史相关性，但我们的系统仍然容易受到由ili相关查询突然增加的情况所引起的错误警报。一个不寻常的事件，比如流行感冒或流感的召回，可能会导致这样的虚假警报。

利用数百万用户的集体智慧，谷歌网络搜索日志可以提供当今最及时、覆盖范围最广泛的流感监测系统之一。虽然传统系统需要1-2周来收集和处理监测数据，但我们的估计是每天进行的。与其他综合征监测系统一样，这些数据是促进进一步的调查和收集疾病活动的直接措施的最有效的手段。

该系统将用于追踪2008-2009年流感季节期间在美国发生的流感类似疾病的传播情况。研究结果可以在网上免费获得http://www.谷歌。org/flutrends.

方法

隐私。在谷歌，我们认识到隐私很重要。我们项目数据库中的任何查询都不能与特定的个体相关联。我们项目的数据库不保留任何关于任何用户的身份、IP地址或特定物理位置的信息。此外，任何超过9个月的原始网络搜索日志都是根据谷歌的隐私政策(http://www.谷歌。com/privacypolicy.html).

搜索查询数据库。就我们的数据库而言，搜索查询是由谷歌搜索用户发布的一个完整的、精确的术语序列；我们不结合语言变体、同义词、跨语言翻译、拼写错误或子序列，尽管我们希望探索这些选项
在未来的工作。例如，我们将搜索查询“流感的迹象”与搜索查询“流感”分开计算
“迹象”和“流感的迹象”。

我们的查询数据库包含5000万个关于所有可能主题的最常见的搜索查询，无需预过滤。数十亿个查询很少发生，并被排除在外。使用与每个搜索查询相关联的互联网协议（IP）地址，通常可以识别出查询来源的一般物理位置，包括在美国境内最近的主要城市。

模型数据。在查询选择过程中，我们使用2003年9月28日至2007年3月11日（包括）期间的所有周拟合每个查询模型，CDC报告ILI百分比，每个区域产生128个训练点（每周是一个数据点）。另外保留了42周的数据（2007年3月18日至2008年5月11日）以进行最终验证。该项目无法使用2003年以前的搜索查询数据。

自动的查询选择过程。使用4倍交叉验证的线性回归，我们将模型拟合到每个区域128个点中的4个96点子集。通过测量模型对32个保留点的估计值与CDC在这些点报告的区域ILI百分比之间的相关性，来验证每个查询模型。时间的

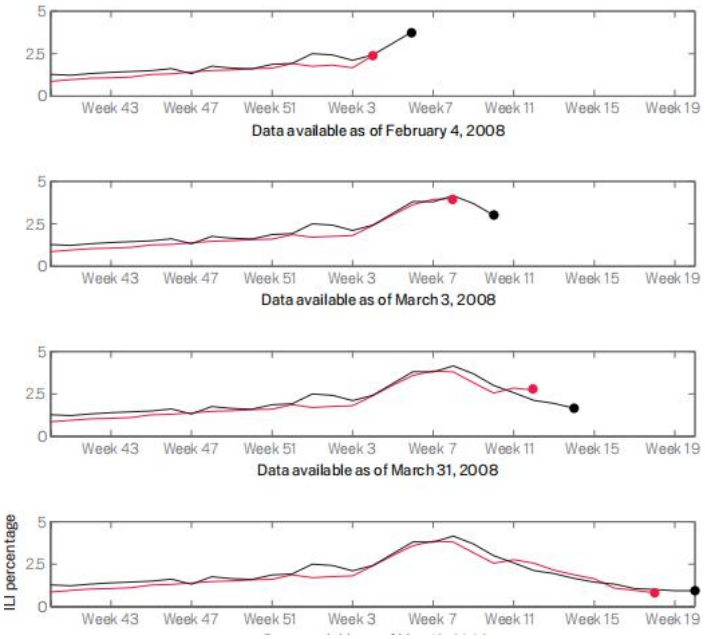

图3：由我们的模型（黑色）估计的和CDC（红色）提供的ILI百分比，显示了2007-2008年流感季节四个点的可用数据。在第5周，我们发现大西洋中部地区的ILI百分比急剧上升；同样，在3月3日，我们的模型显示，ILI百分比在第8周达到了峰值，在第9周和第10周急剧下降。这两个结果后来被CDC ILI数据证实。

我们考虑了滞后时间，但最终没有在我们的建模中使用过程

每个候选搜索查询被评估9次，每个区域一次，使用来自特定区域的搜索数据来解释该区域中的ILI百分比。在每个区域有4个交叉验证折叠，我们获得了候选模型的估计和观察到的ILI百分比之间的36个不同的相关性。为了将这些组合到候选查询性能的单一度量中，我们应用了Fisherz变换[13]，并取36个z变换相关性的平均值。

计算和预过滤。总共，我们拟合了4.5亿个不同的模型来测试每个候选查询。我们使用了一个分布式计算框架[14]以有效地将工作分配给数百台机器。通过假设哪些查询可能与ILI相关，就可以减少所需的计算量。例如，我们本可以在拟合任何模型之前尝试消除与流感相关的查询。
然而，我们担心主动过滤可能会意外地消除有价值的数据。此外，如果得分最高的查询似乎与流感完全无关，那么它将提供我们的查询选择方法是无效的证据。

构造与ili相关的查询分数。我们通过选择保留模型获得最高均值的搜索查询来结束查询选择过程

跨区域的z转换相关性：这些查询被认为是“与ili相关的”。

为了将选定的搜索查询合并到一个聚合变量中，我们在区域基础上对查询分数求和，得到每个区域中与ili相关的查询分数Q的估计。请注意，已为每个区域选择了相同的查询集。

拟合和验证一个最终的模型。我们拟合了最后一个单变量模型，用于基于该区域或状态的与ili相关的查询分数对任何区域或状态进行估计。我们回归了1152个点，结合了9个区域中查询选择过程中使用的所有128个训练点。我们通过测量从最近的时间段（2007年3月18日至2008年5月11日）的42周未测试数据的性能，验证了最终模型的准确性。这42个点约占该项目可用总数据的25%，其中前75%用于查询选择和模型拟合。

状态级模型验证。为了评估使用我们的最终模型生成的州级ILI估计的准确性，我们将我们的估计与犹他州提供的每周ILI百分比进行了比较。由于该模型使用截至2007年3月11日的区域数据拟合，我们使用42周的时间段（2007年3月18日至2008年5月11日），验证了犹他州ILI估计。

致谢我们感谢美国疾病控制与预防中心流感部门的林恩·菲内利对这份手稿的持续支持和评论。我们非常感谢医学博士。犹他州卫生部的罗伯特·罗尔夫斯和丽莎·怀曼以及美国疾病控制与预防中心流感部门的莫妮卡·巴顿提供了ILI数据。我们感谢维克拉姆·萨哈伊对数据收集和处理的贡献，以及谷歌的克雷格·内维尔-曼宁、亚历克斯·罗特和卡塔内·萨尔维安对这份手稿的支持和评论。

作者贡献。J.G.和M。H.M.构思、设计和实现了该系统。J.G.，M.H.M.和R.S.P.分析了研究结果并撰写了论文。L.B.（CDC）提供数据。所有作者都对该论文进行了编辑和评论。

补充材料。数字和其他补充材料可在http://www.nature.com/nature/journal/v457/n7232/补充info/natere07634获得。html

参考文献

1. 世界卫生组织。流感的事实。http://www.谁int/mediacentre/factsheets/2003/fs211/en/ (2003).

2. 世界卫生组织。世卫组织就流感大流行之前和期间的优先公共卫生干预措施进行咨询。http://www.谁int/csr/disease/avian_influenza/consultation/en/ (2004).

3. 弗格森，N。M.以及其他人遏制东南亚新出现的流感大流行的战略。《自然》杂志437、209-214（2005）。

4. 朗吉尼，我。M.以及其他人从源头上含有大流行性流感。科学309,1083-1087（2005）。

5. 埃斯皮诺，霍根，W。&瓦格纳，M。电话分诊：提供监测流感样疾病的及时数据来源。AMIA：年度研讨会论文集215-219（2003）。

6. 马格鲁德，S。评价非处方药的销售可能作为人类疾病的早期预警指标。约翰霍普金斯大学APL技术文摘24,349-353（2003）。

7. 福克斯，S。在线健康搜索2006。皮尤互联网与美国生活项目（2006年）。

8. 赫斯，A.，里德维克，G。和林德。网络查询作为综合征监测的来源。PLoS ONE 4(2)： e4378。doi:10.1371/journal.pone.0004378 (2009).

9. 约翰逊，H。以及其他人对流感监测的网络访问日志进行分析。MEDINFO 1202 - 1206 (2004).

10. 艾森巴赫，G。信息统计学：在网络上追踪与流感相关的综合征监测搜索。AMIA：年度研讨会论文集244-248（2006）。

11. Polgreen, P.M.，陈，Y.，彭诺克，D。M.和福雷斯特，N。D.利用互联网搜索技术进行流感监测。临床传染病病47,1443-1448（2008年）。

12. http://www.cdc.gov/flu/weekly

13. 大卫，F。z和F分布的矩。生物特征36,394-403（1949）。

14. 迪恩，J。和GhemawatS。简化：简化了大型集群上的数据处理。OSDI：第六届操作系统设计与实施研讨会（2004年）。