北京大学互联网金融研究中心
Institute of Internet Finance, Peking University

北京大学互联网金融研究中心工作论文系列

IIF Working Paper Series

NO. IIFWP2015001（总第 1 期）

# 大数据分析的光荣与陷阱

## ——从谷歌流感趋势谈起

沈　艳[1]

2015 年 10 月 27 日

**摘　要：**本文从谷歌流感趋势 2009 年前后表现差异谈起，讨论了大数据分析容易面临的大数据自大、算法演化、看不见的动机导致数据生成机制变化等陷阱，以及对我国大数据产业发展的借鉴。本文认为，为健康发展大数据产业，我国需要防范大数据自大风险、推动大数据产业和小数据产业齐头并进，并强化提高大数据透明度、审慎评估大数据质量等方面的努力。



[1] 沈艳，北京大学互联网金融研究中心高级研究员、北京大学国家发展研究院教授。

# 大数据分析的光荣与陷阱

## ——从谷歌流感趋势谈起

沈 艳

【摘要】本文从谷歌流感趋势 2009 年前后表现差异谈起，讨论了大数据分析容易面临的大数据自大、算法演化、看不见的动机导致数据生成机制变化等陷阱，以及对我国大数据产业发展的借鉴。本文认为，为健康发展大数据产业，我国需要防范大数据自大风险、推动大数据产业和小数据产业齐头并进，并强化提高大数据透明度、审慎评估大数据质量等方面的努力。

## 一、谷歌流感趋势：未卜先知？

“谷歌流感趋势”(Google Flu Trends，GFT)未卜先知的故事，常被看做大数据分析优势的明证。2008 年 11 月谷歌公司启动的 GFT 项目，目标是预测美国疾控中心(CDC)报告的流感发病率。甫一登场，GFT 就亮出十分惊艳的成绩单。2009 年，GFT 团队在《自然》发文报告，只需分析数十亿搜索中 45 个与流感相关的关键词，GFT 就能比 CDC 提前两周预报 2007-2008 季流感的发病率。

也就是说，人们不需要等 CDC 公布根据就诊人数计算出的发病率，就可以提前两周知道未来医院因流感就诊的人数了。有了这两周，人们就可以有充足的时间提前预备，避免中招。多少人可以因为大数据避免不必要的痛苦、麻烦和经济损失啊。

此一时，彼一时。2014 年，Lazer 等学者在《科学》发文报告了 GFT 近年的表现。2009 年，GFT 没有能预测到非季节性流感 A-H1N1；从 2011 年 8 月到 2013 年 8 月的 108 周里，GFT 有 100 周高估了 CDC 报告的流感发病率。高估有多高呢？在 2011-2012 季，GFT 预测的发病率是 CDC 报告值的 1.5 倍多；而到了 2012-2013 季，GFT 流感发病率已经是 CDC 报告值的双倍多了。这样看来，GFT 不就成了那个喊“狼来了”的熊孩子了么？那么不用大数据会如何？作者报告，只用两周前 CDC 的历史数据来预测发病率，其表现也要比 GFT 好很多。

2013 年，谷歌调整了 GFT 的算法，并回应称出现偏差的罪魁祸首是媒体对 GFT 的大幅报道导致人们的搜索行为发生了变化。Lazer 等学者穷追不舍。他们的估算表明，GFT 预测的 2013-2014 季的流感发病率，仍然高达 CDC 报告值的 1.3 倍。并且，前面发现的系统性误差仍然存在，也就是过去犯的错误如今仍然在犯。因为遗漏了某些重要因素，GFT 还是病得不轻。

为什么传说中充满荣光的大数据分析会出现如此大的系统性误差呢？从大数据的收集特征和估计方法的核心，我们可以探究一二。

## 二、新瓶装旧酒：过度拟合

大数据时代的来临，为数据收集带来了深刻变革。海量数据、实时数据、丰富多样的非结构数据，以前所未有的广度进入了人们的生活。但是不变的是，在统计分析方法上，

数据挖掘（Data mining）仍然是统计分析的主要技术。而数据挖掘中最引人注目的过度拟合（overfitting）问题，由于下文提到的各类陷阱的存在，远远没有解决。

我们先用一个故事来解释何为过度拟合。假设有一所叫做象牙塔的警官学校致力于培养抓小偷的警察。该校宣称，在他们学校可以见到所有类型的普通人、也能见到所有类型的小偷；到他们学校来学习就能成为世界上最厉害的警察。但是这所学校有个古怪，就是从不教授犯罪心理学。

象牙塔的教学方式是这样的：将人群随机分为十组，每组都是既有普通人又有小偷。学员可以观察到前九组所有人，也知道谁是普通人谁是小偷。学员要做的是，根据自己从前九组中了解到的小偷特征，从第十组中找出小偷。比如学员从前九组观察到小偷更喜欢在给孩子买尿布的时候也买啤酒，那么在第十组观察到有人在买尿布时也买啤酒，就作为一个嫌疑条件。完成这个过程之后，学校再将人群打散重新分成十组，如此循环往复，之后学校进行测试。测试方式就是再次将人群随机分为十组，看谁能最快最准根据前九组的信息找出第十组的小偷。冠军即象牙塔最棒警察，可以派到社会上抓小偷了。

一段时间后，问题来了：象牙塔最棒警察在象牙塔校内总能迅速找到小偷，可一旦出了象牙塔，该警察就老犯错抓、该抓不抓的错误。他抓小偷的表现，甚至比从来没有来象牙塔学习的人还要差。

在这个故事里，象牙塔最棒警察就相当于根据大数据的数据挖掘方法、机器学习之后挑选出来的最优模型。小偷相当于特定问题需要甄选出的对象，比如得流感的人、不干预就会自杀的人、赖账的人。前九组的人就相当于用于训练模型的训练数据；第十组人则相当于检验训练结果的检验数据。不教授犯罪心理学就意味着抓小偷并不需要理解小偷为什么会成为小偷，类似于在数据分析中只关心相关关系而不关注因果关系。训练最佳警察的过程，就类似于运用机器学习技术，采用训练数据来训练模型，然后采用检验数据来选择模型，并将预测最好的模型作为最佳模型，用于未来的各类应用中。

最后，警察在象牙塔内能快速抓小偷而校外不能，就是过度拟合问题。由于在学校通过多次重复练习，学员小偷的特征已经烂熟于心，因此无论怎么随机分，都能快速找到小偷并且不出错；这就相当于训练模型时，由于已经知道要甄选人群的特征，模型能够对样本内观测值作出很好的拟合。由于象牙塔学校判断小偷的标准主要看外部特征而不去理解内在原因，比如小偷常戴鸭舌帽，那么当社会人群里的小偷特征与象牙塔人群有很大差别时，比如社会上的小偷更常戴礼帽，在象牙塔内一抓一个准的鸭舌帽标准，到社会就变成一抓一个错了。也就是说，在样本内预测很好的模型，到样本外预测很差。这，就是过度拟合的问题。

从过度拟合角度，可以帮助我们理解为什么 GFT 在 2009 年表现好而之后表现差。在 2009 年，GFT 已经可以观察到 2007-2008 年间的全部 CDC 数据，也就是说 GFT 可以清楚知道 CDC 报告的哪里发病率高而哪里发病率低。这样，采用上述训练数据和检验数据寻找最佳模型的方法时标准就很清晰，就是不惜代价高度拟合已经观察到的发病率。Lazer 等人发现，GFT 在预测 2007-2008 年流感流行率时，存在丢掉一些看似古怪的搜索词，而用另外的 5000 万搜索词去拟合 1152 个数据点的情况。

2009 年之后，该模型面对的数据就真正是未知的，这时如果后来的数据特征与 2007-2008 年的数据高度相似，那么 GFT 也该可以高度拟合 CDC 估计值。但现实是无情的，系统性误差的存在，表明 GFT 在一些环节出了较大偏差而不得不面对过度拟合问题。

从上面的故事可以看到，产生过度拟合有三个关键环节。第一，象牙塔学校认定本校知道所有普通人与所有小偷的特征，也就等于知道了社会人群特征。第二，象牙塔学校训练警察，不关心小偷的形成原因，而关注细致掌握已知小偷的特征。第三，象牙塔学校认

为，不论时间如何变化，本校永远能保证掌握的普通人和小偷的行为特征不会发生大规模变动、特别是不会因为本校的训练而发生改变。

在大数据这个新瓶里，如果不避开下面的三个陷阱，就仍然可能装着数据挖掘带来的过度拟合旧酒：大数据自大、算法演化、看不见的动机导致的数据生成机制变化。

## 三、大数据分析的挑战

### （一）陷阱一：“大数据自大”

Lazer 等学者提醒大家关注“大数据自大（big data hubris)”的倾向，即认为自己拥有的数据是总体，因此在分析定位上，大数据将代替科学抽样基础上形成的传统数据(后文称为“小数据”)、而不是作为小数据的补充。

如今，大数据确实使企业或者机构获取每一个客户的信息、构成客户群的总体数据成为可能，那么说企业有这样的数据就不需要关心抽样会有问题吗？

这里的关键是，企业或者机构拥有的这个称为总体的数据，和研究问题关心的总体是否相同。《数据之巅》一书记载了下面这个例子：上世纪三十年代，美国的《文学文摘》有约 240 万读者。如果《文学文摘》要了解这个读者群的性别结构与年龄结构，那么只要财力人力允许，不抽样、直接分析所有这 240 万左右的数据是可行的。但是，如果要预测何人当选 1936 年总统，那么认定“自己的读者群”这个总体和“美国选民”这个总体根本特征完全相同，就会差之毫厘谬以千里了。事实上，《文学杂志》的订户数量虽多，却集中在中上层，并不能代表全体选民。与此相应，盖洛普根据选民的人口特点来确定各类人群在样本中的份额，建立一个 5000 人的样本。在预测下届总统这个问题上，采用这个小数据比采用《文学文摘》的大数据，更准确地把握了民意。

在 GFT 案例中，“GFT 采集的搜索信息”这个总体，和“某流感疫情涉及的人群”这个总体，恐怕不是一个总体。除非这两个总体的生成机制相同，否则用此总体去估计彼总体难免出现偏差。

进一步说，由于某个大数据是否是总体跟研究问题密不可分，在实证分析中，往往需要人们对科学抽样下能够代表总体的小数据有充分认识，才能判断认定单独使用大数据进行研究会不会犯“大数据自大”的错误。

### （二）陷阱二：算法演化

相比于“大数据自大”问题，算法演化问题（algorithm dynamics）就更为复杂、对大数据在实证运用中产生的影响也更为深远。我们还是借一个假想的故事来理解这一点。假定一个研究团队希望通过和尚在朋友圈发布的信息来判断他们对风险的态度，其中和尚遇到老虎的次数是甄别他们是否喜欢冒险的重要指标。观察一段时间后该团队发现，小和尚智空原来遇到老虎的频率大概是一个月一次，但是从半年前开始，智空在朋友圈提及自己遇到老虎的次数大幅增加、甚至每天都会遇到很多只。由于大数据分析不关心因果，研究团队也就不花心思去追究智空为什么忽然遇到那么多老虎，而根据历史数据认定小智空比过去更愿意冒险了。但是研究团队不知道的情况是：过去智空与老和尚同住，半年前智空奉命下山化斋；临行前老和尚交代智空，山下的女人是老虎、遇到了快躲开。在这个故事里，由于老和尚的叮嘱，智空眼里老虎的标准变了。换句话说，同样是老虎数据，半年前老虎观测数量的生成机制，和半年后该数据的生成机制是不同的。要命的是，研究团队对此并不知情。

现实中大数据的采集也会遇到类似问题，因为大数据往往是公司或者企业进行主要经营活动之后被动出现的产物。以谷歌公司为例，其商业模式的主要目标是更快速地为使用者提供准确信息。为了实现这一目标，数据科学家与工程师不断更新谷歌搜索的算法、让使用者

可以通过后续谷歌推荐的相关词快捷地获得有用信息。这一模式在商业上非常必要，但是在数据生成机制方面，却会出现使用者搜索的关键词并非出于使用者本意的现象。

这就产生了两个问题：第一，由于算法规则在不断变化而研究人员对此不知情，今天的数据和明天的数据容易不具备可比性，就像上例中半年前的老虎数据和半年后的老虎数据不可比一样。第二，数据收集过程的性质发生了变化。大数据不再只是被动记录使用者的决策，而是通过算法演化，积极参与到使用者的行为决策中。

在 GFT 案例中，2009 年以后，算法演化导致搜索数据前后不可比，特别是“搜索者键入的关键词完全都是自发决定”这一假定在后期不再成立。这样，用 2009 年建立的模型去预测未来，就无法避免因过度拟合问题而表现较差了。

### (三)、陷阱三：看不见的动机

算法演化问题中，数据生成者的行为变化是无意识的，他们只是被页面引导，点出一个个链接。如果在数据分析中不关心因果关系，那么也就无法处理人们有意识的行为变化影响数据根本特征的问题。这一点，对于数据使用者和对数据收集机构，都一样不可忽略。

除掉人们的行为自发产生系统不知道的变化之外，大数据的评估标准对人们行为的影响尤为值得关注。再以智空为例。假定上文中的小和尚智空发现自己的西瓜信用分远远低于自己好友智能的西瓜信用分。智空很不服气，经过仔细观察，他认为朋友圈言论可能是形成差异的主因。于是他细细研究了智能的朋友圈。他发现，智能从不在朋友圈提及遇到老虎的事，而是常常宣传不杀生、保护环境、贴心灵鸡汤，并定期分享自己化斋时遇到慷慨施主的事。虽然在现实中，他知道智能喜好酒肉穿肠过、也从未见老和尚称赞智能的化斋成果。智空茅塞顿开，从此朋友圈言论风格大变，而不久后他也满意地看到自己的西瓜信用分大幅提高了。

如今，大数据常常倚重的一个优势，是社交媒体的数据大大丰富了各界对于个体的认知。这一看法常常建立在一个隐含假定之上，就是人们在社交媒体分享的信息都是真实的、自发的、不受评级机构和各类评估机构标准影响的。但是，在互联网时代，人们通过互联网学习的能力大大提高。如果人们通过学习评级机构的标准而相应改变社交媒体的信息，就意味着大数据分析的评估标准已经内生于人们生产的数据中，这时，不通过仔细为人们的行为建模，是难以准确抓住的数据生成机制这类的质变的。

从数据生成机构来看，他们对待数据的态度也可能发生微妙的变化。例如，过去社交媒体企业记录保存客户信息的动机仅仅是本公司发展业务需要，算法演化也是单纯为了更好地服务消费者。但随着大数据时代的推进，“数据为王”的特征越来越明显，公司逐渐意识到，自己拥有的数据逐渐成为重要的资产。除了可以在一定程度上给使用者植入广告增加收入之外，还可以在社会上产生更为重要的影响力。这时就不能排除数据生成机构存在为了自身的利益，在一定程度上操纵数据的生成与报告的可能性。比如，在 Facebook 等社交媒体上的民意调查，就有可能对一个国家的政治走向产生影响。而民意调查语言的表述、调查的方式可以影响调查结果，企业在一定程度上就可以根据自身利益来操纵民意了。

简而言之，天真地认为数据使用者和数据生成机构都是无意识生产大数据、忽略了人们行为背后趋利避害的动机的大数据统计分析，可能对于数据特征的快速变化迷惑不解，即便看到模型预测表现差，也难以找到行之有效的克服方法。

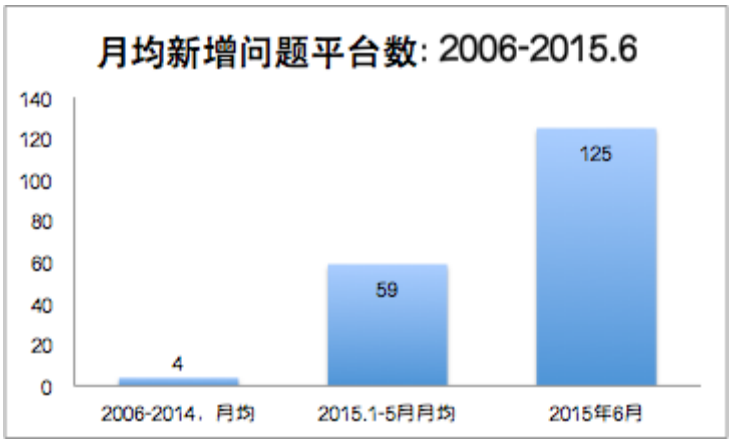
## 四、前车之鉴

目前，我国高度重视大数据发展。2015 年 8 月 31 日，国务院印发《促进大数据发展行动纲要》，系统部署大数据发展工作。《纲要》认为，大数据成为推动经济转型发展的新动力、重塑国家竞争优势的新机遇，和提升政府治理能力的新途径。《纲要》指出，2018 年底前，要建成国家政府数据统一开放平台，率先在信用、交通、医疗等重要领域实现公共数据资源合理适度向社会开放。与此相应，近年来多地成立了大数据管理局、业界学界对于大数据的分析利用也予以热烈回应。因此，了解大数据分析的优势与陷阱，对我国的经济发展和实证研究具有极其重要的意义；而 GFT 项目折射出的大数据使用中可能存在的机会与问题，都值得关注。

### (一) 防范“大数据自大”带来的风险

GFT 案例表明，如果认为大数据可以代替小数据，那么过度拟合问题可以带来巨大的估计误差。这一点在“大众创业、万众创新”的今天尤其需要关注。这是因为大数据作为目前“创新”最闪亮的新元素被高度推崇的，而我国经济处于转型时期的特征，使企业或者机构面对的微观数据不断发生动态变化。如果在数据挖掘中忽略这些变化，往往要面临过度拟合带来的损失。

例如，我国 P2P 网贷行业采用的数据体量虽然大多达不到大数据要求的海量数据，但是不少企业热衷采用爬虫等技术从社交媒体挖掘信息用于甄别客户。这些平台健康状况，就可能与过度拟合的严重程度密不可分。根据中国 P2P 网贷行业 2014 年度运营简报和 2015 年上半年的运营简报，在图一我们可以推算 2006 年到 2004 年间和 2015 年 1-5 月间月均新增问题平台数，并与 2015 年 6 月新增问题平台数作比较。[1]




新增问题平台的大幅增加原因虽然有多方面，但是从数据分析的角度看，由于还没有合法的数据共享机制，P2P 平台在甄别客户质量时，往往只依靠自身渠道和从社交媒体等挖掘的数据，并采用数据挖掘方法建立相应建立模型。在数据分析中，不少 P2P 平台往往疏于查考自身样本的代表性、也忽略宏观经济数据和其他微观数据所包含的信息。由于互联网金融公司出现时间短、又主要成长于经济繁荣期，如果单单依赖有限的数据渠道，数据挖掘与机器学习过程对新常态下个体行为没有足够的认识，在经济下行时仍然根据历史数据而低估逾期率，导致高估平台健康状况，最终不得不面对问题平台不断增加的局面。

### (二) 大数据和小数据齐头并进大势所趋

大数据和小数据各有优劣。简而言之，小数据通常不会假定该数据就是总体，因此收集数据前往往需要确定收集数据的目标、根据该目标设计的问卷或者收集方法、确定抽样框。在数据采集后，不同学者往往可以通过将新收集数据与不同数据的交叉验证，来评估

数据的可信度。小数据在收集上有变量定义清晰、数据生成机制基本可控、检验评估成本相对较低等优点，但是缺点是数据收集成本高，时间间隔长、颗粒度较粗。

大数据的优势就包括数据体量大、收集时间短、数据类型丰富，颗粒度很细。但是，由于大数据往往是一些企业和机构经营活动的附带产品，因此并不是通过精心论证的测度工具生成。另外，由于大数据的体量很大，交叉验证数据的可信度、不同学者采用相同数据独立研究以检验数据的前后一致性等工作难度较大。这些特点意味着大数据本身未必有科学研究要求的那样准确、可靠，在数据分析中就需要对大数据适合研究的问题有较清晰的认识。

在与小数据互为补充推动研究与认知方面，大数据大有可为。将大数据与小数据相结合，可以大大提高数据的颗粒度和预测精度。比如对 CDC 流感发病率的预测研究发现，将 GFT 采用的大数据和 CDC 的历史数据相结合的模型，其预测能力比单独运用大数据或者小数据要好很多。

大数据往往可以实时生成，对于观察特定社区的动态具有小数据无可替代的优势。比如，美国在“九一一”之后，出于快速准确估计在某个特定小社区活动的人口的需要而启动了“工作单位和家庭住址纵向动态(LEHD)”项目,该项目将人口普查数据、全国公司数据、个人申请失业保险、补贴、纳税等记录联通，可以对社区在短时间内的“新陈代谢”作出较为全面的刻画。

这类的数据结合研究，对于了解我国社会经济状况的动态变化会十分重要。一个可能的应用是，将城市人口、工作状态、性别、年龄、收入等小数据采集的信息，和实时产生的交通状况相结合，来预测人们的出行特征，来解决城市交通拥堵、治理雾霾等问题。另一个可能的应用是，推动人民银行征信中心个人征信系统数据和民间征信系统大数据的结合，建立高质量的中国个人征信体系。

另外，我国经济处于转型时期，有不少政策亟需快速评估政策果效。以小数据为基础，利用大数据数据量丰富的优势，可以通过互联网做一些随机实验，来评估一些政策的效果，也是可能的发展方向。

在过去的十多年中，我国在通过非官方渠道采集小数据、特别是微观实证数据方面取得了长足进展。在多方努力下，更多经过严格科学论证而产生的数据可被公众免费获得并用于研究。例如，北京大学的“中国健康与养老追踪调查”、“中国家庭追踪调查”，都由经济、教育、健康、社会学等多领域的专家协同参与问卷的设计和数据采集的质控。在这些努力下，小数据的生成机制更为透明，交叉验证调查数据的可信度等实证研究的必要步骤也更为可行。

但是，目前在小数据的收集和使用、政府和有关机构的小数据开放运用方面，我国还有很大推进空间。只有在对涉及我国基本国情的小数据进行充分学习研究之后，我国学界和业界才能对经济政治社会文化等领域的基本状况有较清晰的把握。而这类的把握，是评估大数据质量、大数据可研究问题的关键，对推进大数据产业健康发展有举足轻重的作用。

因此在政策导向上，为要实现大数据、小数据相得益彰推动经济发展的目标，在促进发展大数据的同时也要大力发展小数据相关产业，推动小数据相关研究与合作，使大数据与小数据齐头并进、互为补充。

### （三）提高大数据使用的透明度，加强对大数据质量的评估

大数据面临的透明度问题远比小数据严重。在 GFT 案例中，Lazer 等人指出，谷歌公司从未明确用于搜索的 45 个关键词是哪些；虽然谷歌工程师在 2013 年调整了数据算法，

但是谷歌并没有公开相应数据、也没有解释这类数据是如何搜集的。我国大数据相关企业的数据，也鲜有学者可以获得并用于做研究的例子。

与透明度相关的就是大数据分析结果的可复制性问题。由于谷歌以外的研究人员难以获得 GFT 使用的数据，因此就难以复制、评估采用该数据分析结果的可靠性。因此利用大数据的研究难以形成合力，只能处于案例、个例的状态。

另外还要注意到，如果数据生成机制不清晰，研究结论难以复制，而算法演化也表明，最终数据往往成为使用者和设计者共同作用的结果。这种数据生成的“黑箱”特征，容易成为企业或者机构操纵数据生成过程和研究报告结果的温床。唯有通过推动大数据的透明化、公开化，我们才能在大数据产业发展之初，建立健康的数据文化。

因此，在大数据时代，为了更好利用大数据，需要采取相关措施，增加在大数据生成过程的透明度方面的努力。例如，采取措施推进数据生成企业在妥善处理隐私信息后，定期公布大数据随机抽样数据、要求数据生成企业及时公布数据算法的变更，鼓励采用大数据的研究实现可复制性、便于交叉验证等。

## 五、结语

目前有些流行观点认为，在大数据时代，技术容许人们拥有了总体因此抽样不再重要、另外由于数据挖掘术的进展，只需关心相关关系而不必再关心因果关系。而 GFT 的实例表明，即便谷歌公司用于 GFT 计算的是数十亿的观测值，也不能认为谷歌公司拥有了流感人群的总体。误认为数据体量大就拥有了总体，就无法谦卑结合其他渠道的小数据，得到更为稳健的分析结论。而 GFT 估计的偏误原因，从来都离不开人们的主动的行为-- 无论是谷歌公司自己认为的 GFT 的流行导致更多人使用该搜索、还是 Lazer 等人认为的算法变化、丢弃异常值。因此，不明白数据生成机理变化的原因而只看相关关系的后果，于谷歌是 GFT 的计算偏误丢了脸，而对热情地投身于采用大数据到创新、创业中的中国民众和相关机构来说，则可能是不得不面对事先没有预备的重大经济损失。

[1] 《2015 年上半年 P2P 网贷简报：新上线平台数接近 900 家》
http://p2p.hexun.com/2015-07-01/177192976.html