打开共享社会媒体数据之门*

作者：弗雷德·莫斯得特(Fred Morstatter)
刘　欢(Huan Liu)
曾大军(Daniel Zeng)
译者：文益民　蔡国永　闭应洲

关键词：社会计算　共享数据

自科学发现之初，科学家和研究者就一直在系统地记录他们的数据，这些记录下来的数据对于知识进步一直非常重要。最近，主要是由于互联网 (the Internet)，这些数据呈现出指数增长势头[1]。世界各地的数不胜数的研究小组已经提出并实现了很多掌握这些数据的方法，并试图让所有人访问这些数据。如美国加州大学欧文分校的数据仓库[1](data repositories, http://archive.ics.uci.edu/ml）已经被证明非常成功，它组织并存储的数据集和基准在其他科学家的研究工作中得到了使用。数十年来，研究者们因使用这些数据存储而减少了他们花在数据获取上的时间。这帮助了科学家们以更快的速度生成知识。

共享数据对于研究者研究他们所关注的领域有诸多好处。首先，它鼓励了该领域的业余爱好者。例如，一个关于乳腺癌的公共数据集不仅能帮助肿瘤学家，还能帮助那些缺乏仪器设备而无法自己生成这些数据的刚入门的癌症研究者。另一方面，共享数据可供全球范围内的学术出版物的作者和评阅者访问，在数据的特性和表现已经为学术界所熟悉的情况下，有助于防止涉嫌学术不端论文的发表。当一个研究者引用一个数据集时，读者能获取该数据集来独立地对该作者的工作进行验证。

基于此，研究者们正开始在发布整个数据集的同时还发布相应的软件，使未来的评阅者和评价者能够使用这些数据和软件重现作者的研究结果。用这种方式发布的研究工作被称为可重现研究，它曾被谷歌流感趋势预测所使用（这是2009年一个研究工作的副产品，它的可重现部分位于www.google.org/flutrends[2]）。这种方法及与产生的所有结果一起发布的用于分析的工具代码正引起研究界的关注[3]。一位科学家甚至还创建了一个计算尺来度量研究工作的可重现性，并且附加在他作为编辑的刊物上[4]。将代码和数据包含在一起能够体现研究者用来得到发表在刊物中的结论的逻辑。数据分发和可重现研究给评阅者提供了可靠可信的方式获取数据以检验作者的工作，使得出版物增加了可信度。

大约十年以前，一种新出现的媒体吸引了所有领域的研究者的注意力——“社会媒体”或者“社交网络”已经从一个含义模糊的术语发展成为互联网上几乎每个人日常生活的一部分[5]。诸如Facebook[2]和LinkedIn这样超大规模的在线社区已经吸引了世界各地的数据挖掘者、心理学

* 本文译自*IEEE Intelligent Systems* 2012年第1期的Opening Doors to Sharing Social Media Data一文。

[1] 亦有译作“数据仓储”。

[2] 脸书，也有译作“脸谱”。

家、社会学家以及其他社会科学家们的研究兴趣。尽管这个新的研究领域给研究者们提出了一些引人入胜的问题，但这些社交网络站点却拒绝向公众发布他们的数据，同时也禁止有兴趣的第三方采集他们的数据（例如，Foursquare的服务条款中明确地申明，“不能直接或者间接地采集或者截取网站的任何内容”）。研究者还不得不处理令人头痛的事情，即社会媒体网站会追查谁发布了他们的数据。比如Twitter(推特）要求美国斯坦福大学移除他们用于研究的数据（这个数据集以前发布在：http://snap.stanford.edu/data/twitter7.html）。

本文结合当前环境概述社会媒体数据共享中产生的问题，描述一种同时支持开放性标准及对数据集的分布式访问的产生和传输数据集的方法。最后，我们试图在科学界发起一次关于如何应对数据共享困难的对话。

封锁数据的危险

数据开放对于科学合作非常关键。没有有效的数据交流，研究工作就缺乏可信度；没有适当的数据支持，科学结果也会失之交臂。共享数据的开放会有助于保全科学的声誉和研究结果的可信度，虽然不道德的科学家并不多。更重要的是，封锁数据会从根本上产生不良影响。当科学家们不肯分享或被禁止分享数据时，由于科学家不能从其他研究者采集的数据受益，所有的科学都会受到影响。尽管从表面上看来这显得无关紧要，但深入思考后，我们能看到这会对优秀的科学研究造成威胁。很多领域里的研究工作依赖于相同的数据，以确保研究结果匹配或者能完整地检测一个方法是否正确。比如，在特征选择领域，美国加州大学欧文分校(UCI)机器学习库中的乳癌数据集(http://archive.ics.uci.edu/ml/datasets/Breast+Cancer)给了科学家们一个共同的基础去测试他们的算法。UCI网络数据集上的空手道俱乐部数据集(Zachary’s Karate Club)[6]对于社会计算研究也扮演了同样的角色。假如这些数据突然被移除，那么由于实验数据不再对公众开放，成千上万的论文会在一定程度上丧失其可信度。

我们提出的共享数据的方法

尽管已经有人在努力解决数据封锁带来的问题，比如美国亚利桑那州立大学建立了社会计算的中心数据仓库，但由于无法说服社会媒体网站允许数据仓库共享网站的数据，结果无法生存。无能为力改变这种状况给当前科学研究的工作流程带来了严重漏洞，这种状况必须得到修补，以确保今天的科学研究在今后几年仍能经得起考验。当关于共享数据的争论还在延续的时候，研究者们正开始接受一种观点——如果研究者继续按照传统的方式存储他们的数据（数据存储于在线数据仓库里），那么这些数据可能无法被评阅者无限地访问或者为未来的研究工作所使用。只要这种状况存在，我们就必须提供一种非中心的工具，使得中心数据仓库不存在时，数据仍然可被用户访问。这里描述我们提出的一种方法，该方法使拥有技术想法的读者能有效地访问他所需要的数据。我们的研究背景是社会计算，并且我们的方法主要偏重于社会计算的应用。

为了克服很多网站给数据共享和数据分发强加的严格规则，我们开发了一种依赖于研究中所采集数据的参数而不是数据本身的方法。我们使用的这些参数只关注数据的整体网络结构，其原因是:第一，参数足够体现原型；第二，很多研究工作仅使用这种类型的数据；第三几乎所有的社交网络站点都提供这种数据。这些参数所描述的模型易于应用到几乎任何社会媒体输出的数据。我们为数据采集选择了以下的参数向量：<网站名字，初始的网络用户，需要爬取的网络用户数量>。

我们从20多个社会媒体网站采集了数据来测试我们的方法。首先，我们为每个网站编写了一个爬虫软件，这些网站在提供给其一个适当的初始网络用户后，可以从中得到HTML响应，从而可以抽取网页上必要的信息（在本文中，我们定义的爬虫

软件是指专为从一个特定站点的HTML文档中抽取信息的可执行程序）。需要读者注意的是，我们只关注社交网络站点上的网络数据，我们为每个爬虫软件编写了以下函数：getFriends, getFollowers和siteName。前两个函数使用了同一个参数，即我们希望抓取的社交网络站点的用户，我们称之为**目标用户**。在有向网络中，getFriends旨在获取目标用户的所有朋友；getFollowers旨在获取跟随目标用户或者订阅目标用户（博客）的全部用户（对于无向图，getFriends与getFollowers这两个函数产生的结果相同）；最后，siteName函数返回的是与正在抓取的站点有关的静态数据。在后面的步骤中，以上程序片段将被组织成一个完整的应用程序。

在为每个网站构造一个单独的程序片段后，我们构建了一个总程序，以一种友好、直接的方式将程序片段打包成爬虫软件交给研究者。这项工作的目的不仅是为了让我们的软件好用，还为了方便没有计算机编程基础的社会科学家们使用我们的工具。为了实现这个想法，我们设计了HTML图形用户界面程序，让研究者输入他们需要的参数。我们的程序能通过异步信息传递一次执行多项任务，一旦选择的爬虫软件运行结束，用户就可以把数据存储到本地。图1描述了该系统是如何工作的。

一旦使用我们系统的用户完成了他们的实验，他们只需要简单地将他们在爬取数据时使用的参数作为文本加在他们的论文里就行了。使用这样的方法，评阅者或者未来的读者能够通过使用上述软件的本地拷贝重新获取实验数据，从而对论文的结果进行验证。下载的数据并没有向公众发布，而评阅者和研究者都能知道数据是如何构造出来的。

美国国家标准与技术研究院已经使用了类似的方法来发布文本检索会议(TREC)的Tweets2011数据集，它包括一千五百万条微博数据(http://trec.nist.gov/data/tweets/)。他们的方法依赖于Twitter的一个特殊例外，即人们通过Twitter可以共享在数据采集中使用的微博ID号。TREC不但公布了ID列表，也公布了下载微博的爬虫程序。我们的方法在两个重要方面与他们的方法不同：第一，我们的方法对任何社会媒体站点都有效；第二，我们的方法关注的是

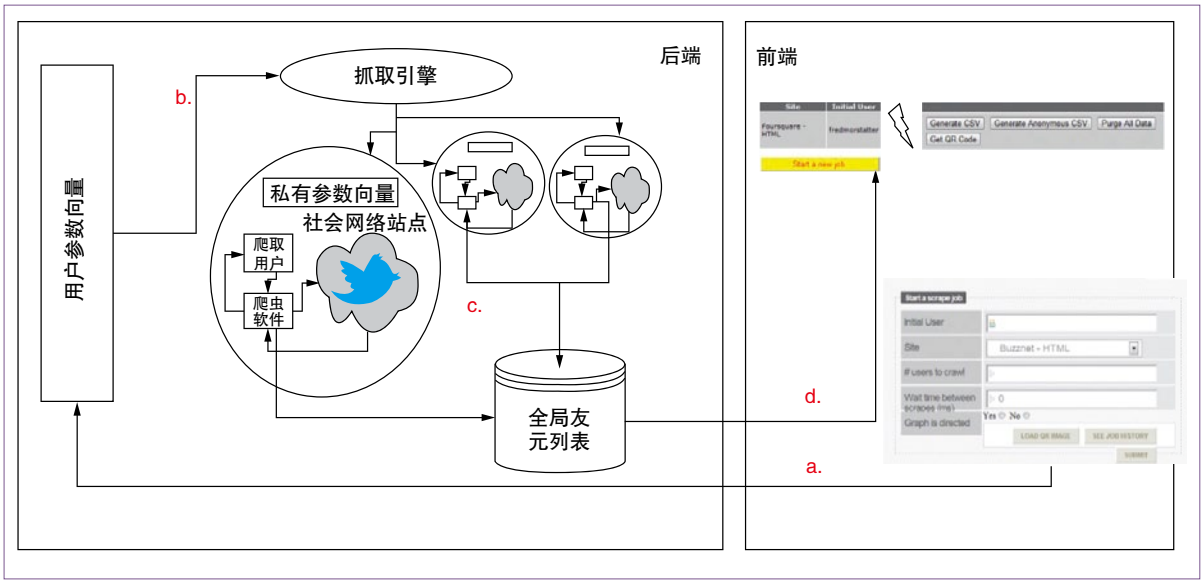

图1 一个爬虫软件原型的高层视图：(a)图形用户界面从研究者那里获得参数并将其转化成参数向量；(b)参数向量被送入消息队列，当需要的资源就绪时，它就会被使用；(c)爬虫软件爬取参数向量中要求的每个页面的信息，并将每次调用获得的结果存储在一个总表中；(d)爬虫软件操作结束后，用户可以下载数据

表1 使用参数向量<Foursquare, fredmorstatter, 500>在两个不同时间获取网络的简单特征

特性	2011年2月23日	2011年12月7日
节点的数量	5628	8297
边的数量	8857	14855
网络直径	8	4
每个节点的平均邻居数量	3.119	3.516

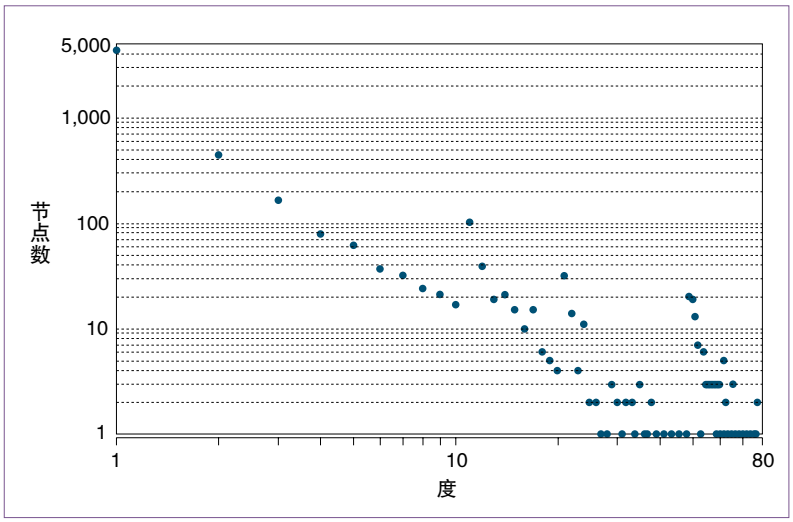

图2 2011年2月23日获得网络的度分布。在此图中我们看到一些节点的连通性遵循幂律分布

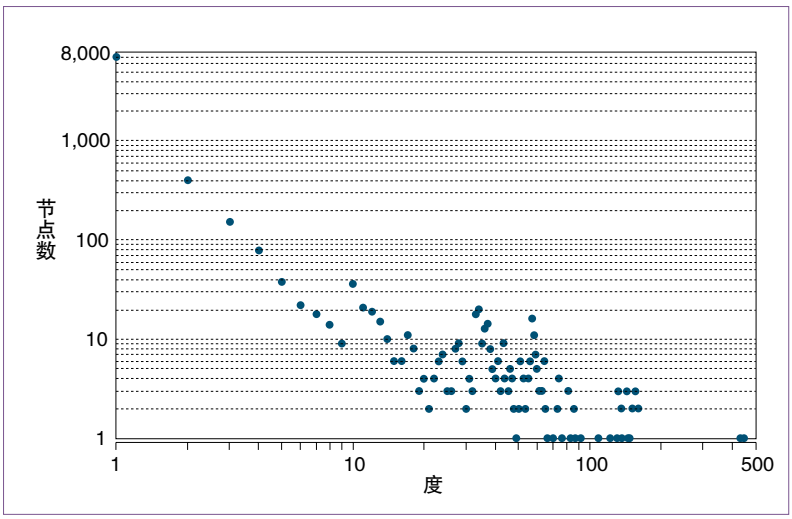

图3 2011年12月7日收集使用相同的参数向量获得网络的度分布。我们仍然可以看到相同的分布特征，但节点更多

网络结构信息而非网站的内容。

评测我们的方法

我们描述的数据发布机制为研究者传送他们研究工作中的数据提供了一种新的方式。然而，我们还关注这个方法的稳定性，这是因为在相同参数下采集的两个数据集，如果采集的时间相隔较长，呈现出的特性肯定不同。我们通过从 Foursquare 网站上下载了初始用户为“fredmorstatter”的 500 条数据以研究这种预期的变化，使用的参数向量是 <Foursquare, fredmorstatter, 500>。我们将此任务执行了两次，一次是在 2011 年 2 月 23 日，另一次是在 2011 年 12 月 7 日。表 1 体现了两个数据集的特性，图 2 和图 3 则展示了不同的度分布。尽管从两个数据集中发现了明显的不同，我们也能观察到它们是在按照一种合理的方式演化。例如，在差不多 9 个月的时间里，网络的直径减小了一半。这种网络直径的收缩在其他网络里也能观察到 [7]，而且这种相同的结果也表明我们的数据采集策略看来是正确的。

局限性

为每个读者仓促地创建一个新的数据集有内在的局限性。首先就是数据的可靠性问题，因为数据会随时间而演化。可以肯定的是，在研究者刚开始采集数据

时，一个拥有150个朋友的用户在3个月之后的论文送审时，可能不再拥有那么多朋友。用户如何确保使用这种方法得到的数据集的性质能得到保持？这对于规模小的数据集来讲更是个大问题，因为一个小规模的用户集在很短的时间里可能会产生变化。对于规模大的数据集，我们提出的方法可能会表现得很好，这是因为在规模大的网络中新增一条边，不会影响到数据集的整体性质。

可是，在我们提出的方法中使用规模大的数据集会带来另一个问题：在理想状态下，获取规模大的数据集会给社交网络站点带来显著的流量。这个问题不可小看，因为很多社交网络站点禁止对它们的服务器产生这种流量，并且常常会追查是谁违反了它们的规则[8]。结果，研究人员可能会因产生的流量被网站拥有者起诉。为此，在爬取网站内容之前，弄清楚网站的服务条款，并获得网站拥有者的允许往往是很重要的。另外，尽管网站可能批准以研究为目的的爬虫的使用，但如果不怀好意的垃圾邮件制造者也使用该爬虫软件进行拒绝服务攻击，软件的作者还得为他们造成的任何损失负责。

数据共享对于任何科学探索都是至关重要的，但是社会媒体和数据挖掘领域正受到社会媒体站点施加的数据获取限制所带来的严重挑战。尽管我们提出的工具意在减轻社会媒体站点这些限制所带来的压力，但并不意味着这是一种“灵丹妙药”，而只是打开了关于数据共享的大讨论之门。作为科学家，我们需要设计一种能实现数据共享的协议，这种协议既能保护我们所研究的用户的隐私，又能遵守社交网站的规定。我们诚邀读者进一步拓展我们提出的方法，或者提出完全不同的思路，以帮助科学家们解决这个领域内最严重的问题之一。■

作　者：

弗雷德·莫斯得特(Fred Morstatter)：美国亚利桑那州立大学在读博士、亚利桑那州立大学数据挖掘与机器学习实验室研究助理。fred.morstatter@asu.edu

刘欢(Huan Liu)：美国亚利桑那州立大学教授。AAAI、ACM会员，IEEE会士。huan.liu@asu.edu

曾大军(Daniel Zeng)：复杂系统管理与控制国家重点实验室和北京市智能化技术与系统工程技术研究中心教授、美国亚利桑那大学教授。zeng@email.arizona.edu

译　者：

文益民

CCF高级会员。桂林电子科技大学教授。主要研究方向为机器学习与数据挖掘等。ymwen2004@yahoo.com.cn

蔡国永

CCF高级会员。桂林电子科技大学教授。主要研究方向为社会网络及社会媒体数据处理等。ccgycai@guet.edu.cn

闭应洲

广西师范学院教授。主要研究方向为智能计算、智能信息处理。yingzhou.bi@gmail.com

参考文献

[1] K. Coffman and A. Odlyzko, The Size and Growth Rate of the Internet, First Monday, vol. 3, no. 10, 1998; http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/620/541

[2] J. Ginsberg et al., Detecting Influenza Epidemics Using Search Engine Query Data, Nature, vol. 457, no. 7232,2009, 1012~1014

[3] F. Leisch, Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis, Proc. Computational Statistics (Compstat 2002), W. Härdleand B. Röz, eds., Physica-Verlag, 2002

[4] R. Peng, Reproducible Research in Computational Science, Science, vol. 334, no. 6060, 2011, pp. 1226~1227

[5] D. Zeng et al., Social Media Analytics and Intelligence, IEEE Intelligent Systems,vol. 25, no. 6, 2010, 13~16

[6] W.W. Zachary, An Information Flow Model for Conflict and Fission in Small Groups, J. Anthropological Research,vol. 33, no. 4, 1977, 452~473

[7] J. Leskovec, J Kleinberg, and C. Faloutsos,Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations, (KDD 05), ACM, 2005, 177~187

[8] D. Heath, Pete Warden vs. Facebook: A Case of Too Much Data Access,ITWire, 7 Apr. 2010; www.itwire.com/it-policy-news/regulation/38101-petewarden-vs-facebook-a-case-of-toomuch-data-access