大数据与失业分析*

米哈埃拉·西米欧奈斯库 克劳斯·F·兹姆曼

[摘要] 互联网数据或者说“大”数据，正被越来越多地用来及时评估个人、家庭、企业和公共机构的相关活动。信息集涵盖大量的观测值，并可以容纳灵活的概念形态及实验设置。因此，对于研究广泛的人力资源议题，如预报、现报和确定卫生及福利问题，把握个人生活方方面面的匹配过程，以及在使用传统数据评估存在缺陷的复杂问题中，互联网数据极其有用。对于多数国家而言，互联网数据能够改进失业模型和失业预测分析。不过，预测的准确度还要取决于一个国家的互联网普及率、互联网用户的年龄结构，以及所建构的互联网变量的稳定性。

[关键词] 大数据；失业；互联网；谷歌；互联网普及率

[作者简介] 米哈埃拉·西米欧奈斯库：罗马尼亚科学院经济预测研究所高级研究员，布拉格商学院移民研究中心教授，全球劳动研究中心研究员；克劳斯·F·兹姆曼：普林斯顿大学客座教授，荷兰联合国大学马斯特里赫特技术与创新研究所人口发展与劳动经济中心主任，全球劳动研究中心主任

一、引言

互联网数据，特别是谷歌网站上的检索行为数据，已被不同领域的研究者用于对不同变量的即时预报、预测或分析。例如，对经济学家和决策者来说，及时了解宏观经济指标的真实状况至关重要。然而，在多数情况下，这些重要信息只能由国家统计部门发布，不仅具有一定的滞后性，有时还被校正过。2008 年末爆发经济危机期间，关于经济受创程度的官方数据不能提供有效信息。相比之下，互联网数据不仅能做出即时预报，而且还可以为分析个人、企业和机构行为提供潜在有效的数据。

本文旨在考察互联网搜索数据在各个领域，特别是在对不同国家的失业状况进行建模方面是否有用。针对一些发达国家的经验研究已经证实了大数据对于失业率预测及建模的有效性。

数字革命标志着从模拟和机械电子技术向数字电子技术的演变，代表着信息时代的到来。数字逻辑电路及关联技术（互联网、计算机、数字移动电话）的大量生产和广泛应用是数字革命的主要支柱。为了建成一个数字驱动经济和数字驱动社会，网络计算正越来越多地融入我们的日常生活。[1]

生活的方方面面均可记录在案。个人和企业的一举一动都在互联网上呈现，这能够反映市场经济的完整画面，以及嵌入大数据云里的社会生

* 本文中文翻译：中央财经大学中国互联网经济研究院史珍珍、中国人民大学学术期刊社李存娜；校对：史珍珍。

活全景。意识到这些数据蕴含着巨大研究潜力的社会科学家可以利用这些信息资源。历史数据可以重复分析，以便不断更新对某种现象或进程的看法。利用互联网，在问题提出之前就能给出答案，这意味着研究者可以考虑新的研究策略和新颖的调查设计。

各种产品和服务的在线市场发展迅速，其中受到特别关注的是就业市场。社交媒体喜欢这一现象，它们拥有关于个人行为及偏好的大量数据。[2]由于技术嵌入了日常生活，社会成分正朝新的方向飞快发展。数字技术、信息经济学和通信技术的最新进展，显示了宏观与微观意义上的“第二经济”（second economy）的重要性。[3]第二经济是数字时代的核心，它在物质世界中安置了一个神经系统。在美国，第二经济的规模将很快超过实体经济。

第二经济中利用率最高的部分是互联网，社交媒体就是在互联网上运作的。时下有许多非常流行的产品，像 Google＋，Facebook，LinkedIn，Ywitter 或 YouTube。来自第二经济、微型化技术、社交媒体和互联网的数据，能够对官方统计数据形成补充。[4]对于那些研究者感兴趣的、被大量访问的关键词，通过谷歌，就可以获得关于它们的实时、高频、集总数据。[5]但是，谷歌并没有对这些数据的阈值做出说明。

尽管经济学研究方法取得了长足的进步，但它在测量上仍存在缺陷，许多指标要么是刚确定下来，要么是经常修改。在此背景下，互联网检索数据即使存在局限性，也仍然代表着一个有意义的替代选项，具有巨大的潜力。对分析和预测失业而言，谷歌检索数据非常有用。

在本文第二部分，我们会从总体上讨论一下互联网活跃数据。第三部分则集中探讨用于失业率建模的互联网数据，此处将从讨论阿斯吉塔斯（N. Askitas）和兹姆曼（K. F. Zimmermann）的宝贵经验[6]开始。第四部分为本文结论。

二、互联网活跃数据

在 20 世纪 80 年代互联网兴起之时，社会科学研究者认为互联网为通过在线调查或其他方法收集数据提供了良好的环境，其优势在于价格和速度。到了 90 年代，互联网蓬勃发展，成为人们日常生活的一部分，因为它有这样的好处：人们可以通过电子邮件和其他设备进行快速的沟通，可以上网冲浪或搜索特定的答案。[7]进入 21 世纪以来，在网络技术发展的同时，各种技巧也更加完善。人们对互联网的使用越来越多，互联网产生了大量的数据。一开始，人们甚至不知道他们的数据被收集和存储了。传统调查在收集数据时需要取得调查对象的同意，与此不同的是，现在人们在家庭隐私空间或在办公室的行为和偏好都被观察研究了。随着谷歌进军市场领域，各种个人信息都被传播出去。除谷歌之外，流行的互联网数据源还有 Facebook，Wikipedia，LinkedIn，Twitter 等。

康斯坦特（A. Constant）、兹姆曼[8]及阿斯吉塔斯和兹姆曼[9]最早发表了研究谷歌活跃数据是否对分析社会议题有用的研究论文，这些议题包括美国总统大选、失业、经济大衰退等。戈艾尔（S. Goel）等人针对互联网活跃数据做了一个大型调查，描述了大数据的强项和弱点所在。[10]互联网数据有很多优点：它们是数字生成的，便于存储、组织和处理。它们有地理标记和时间戳，可以进行横截面的与横向的精准测量。[11]运用互联网数据，能够为提高社会福祉做出更加明智、及时、有效的决策，特别是在危机时期。在此背景下，理论与经验数据的关系就改变了。大数据涉及大量的观测值，允许采取灵活的实验设置和概念形态。搜索活动数据使研究者能在不同的时空背景下进行分析，有利于跨学科研究，并能提供间接的面板调查数据。在经济危机期间，由于互联网数据是以高频率且近乎实时提供的，因此经济受破坏的趋势能够及时被察觉。

互联网数据的弱点，可能与它们只能以集总数据的形式提供有关。[12]其方法如何，没有完备的记录。互联网活动是通过选定的搜索关键词来反映的，然而，这些关键词合适与否，可能随着区域和时间段的不同而有差异。谷歌网页排名会影响供求。地理位置是用 IP 地址来界定的，但这些地址只在国家层面才能获得。一些小的领域还需改进。此外，因为互联网的使用可能有偏向性，那么即便样本是基于大量互联网活动得出

的，样本也未必能代表整个群体。例如，麦克拉伦（N. McLaren）和珊布格（R. Shanbhogue）的研究表明，互联网的使用会因收入和年龄的差距而有所不同。[13]

鉴于不同个体、不同国家应对新技术浪潮的方式不同，选择性偏差是个重要问题。[14]互联网普及率是指一个国家全部人口中互联网用户所占的比例。有的国家互联网普及率高达 90%以上，但在另外一些国家这个比率要小得多。2016 年 6 月 30 日更新的欧盟互联网统计数据显示，欧盟的互联网普及率为 80.1%。2016 年，德国的互联网普及率是 89%，英国是 91.6%，丹麦是 95.9%，挪威是 96.3%，而美国仅为 88.1%。[15]即使在互联网高度普及的国家，也不是人人都使用社交媒体或智能手机，而这会导致选择性偏差。

将来人们会越来越多地通过（客观的）嵌入式传感器来获取新的数据，这能提供关于个人生命体征、位置、人类活动与经济活动的信息。如此，我们的经济会越来越依赖数据，而研究机遇也会增加。就像阿斯吉塔斯和兹姆曼指出的那样，新技术及其组合将会产生新的数据并带来新的挑战。[16]

调查者在样本容量、样本规模、采样频率上存在的地理差异因互联网数据得以弥补，而且使用在线调查或电子邮件不会产生边际成本。[17]作为一个调查平台，互联网既带来了方法论上的新挑战，同时也具有巨大的潜力。由于互联网无处不在，所以既可以获得代表性样本，也可以获得随机样本。在充分占有数据的情况下，选择性偏差就被消除了。因为在线用户的特征非常接近于总人口，因此，样本就有了代表性，而且还是随机的。这样一来，由于拥有无限的数据，抽样就不再是必需的了。在根据互联网数据进行的大规模调查方面，一个著名范例是工资指标基金会（Wage Indicator Foundation）进行的工资指标调查。① 基于个人报告形成的工资调查有 20 多种语言的版本，涉及 60 多个国家。统一化的工资数据对大量的样本国开放。选择性偏差的问题虽然存在，但进一步的研究正在试图弥补这个缺陷。

用互联网进行调查成为数据收集的重要渠道。信息和通信技术与互联网的优势在于，它们能够减少几乎所有市场上匹配工作中的搜索摩擦。匹配不仅在现实生活中极其重要，对于经济学来说也是如此，因为匹配问题及最优解是其研究对象和目标，例如，将长途旅客和飞机的座位相匹配，或者把游客和出租车相匹配。其他的例子还有在就业市场[18]和婚姻市场[19]上对个人进行配对，这也凸显了互联网在减少搜索摩擦上的优势。而这还可以带来新的商业机会，例如招聘服务和网上相亲服务。这种针对不同背景下经济行为的新的数据潜力，有利于我们富有成效地重新思考那些久拖未决的问题。实际上，互联网还使不同的劳动力市场被取代。例如，如果有人需要医生、律师和装配工等的帮助，他/她只要输入相应的关键词，就能在很短的时间内从网络上得到数百个选项。另外，许多雇主用互联网（比如通过 LinkedIn）来招聘雇员。2008—2010 年的经济大衰退也证实了互联网的巨大潜力，因为此时人们都集中到网上去找工作了。

互联网搜索引擎市场将文献的供给与需求进行了匹配。将信息的需求与包含此类信息的文献的供给相关联。因此，互联网可以及时反映信息需求的整体状况，而我们也能就此了解检索此类信息的个人的状况。Google Trends 和谷歌的商业模式为我们展现了这一需求的全球图景。阿斯吉塔斯和兹姆曼的研究就遵循了这一思路，突出了 Google Trends 数据的应用，而阿斯吉塔斯和兹姆曼则强调了技术数据的应用。[20]

Google Trends 这一数据供应工具在 2008 年夏天开始投入使用，目的是公布对某些问题的相对网络检索量，其中用户可以自由界定针对这些问题的关键词。Google Trends 会根据特定地区的用户在谷歌上查询的问题的多少，给出一个时间序列指数。这个查询指数的计算方式为：某个地区对特定关键词的查询总量，除以该地区某个时间段内查询问题的总量。该时期最大的查询份额规定为 100，起始阶段的查询份额为 0。[21]

阿斯吉塔斯对 Google Trends 的优点和缺陷

① 参见 http://www.wageindicator.org/main/Wageindicatorfoundation/researchlab/wageindicator-survey-and-data.

有精当的描述。[22] Google Trends 团队用“会话分析”（sessionization）这一术语来表示搜索数据都经过了标准化处理，减少了由于打字错误、草率的重复、改写和其他行为导致的数据噪音。搜索会话可以分布在基于 IP 地址——会话正肇始于这些 IP 地址——的不同地区。其科学潜力在于，用户有能力界定相关变量集，并通过界定及合并关键词建构搜索内容。因此，我们有可能轻轻松松地检视不同概念带来的不同结果。

然而，Google Trends 这种工具只能让我们得到关于微观行为数据的总体印象。它对其方法没有很好的描述，也没有命名版本号。对于大规模的检索数据、互联网高度普及的地方，Google Trends 是有效的。但 IP 地址只在国家层面才能获得。而且，对数据的获取也受到谷歌的制约，因为它可能改变关于数据供应的承诺。同时需要指出的是，当新的数据组创建时，Google Trends 提供的数据是从新近抽取的代表性子样本获得的。因此，研究者需要存储数据，以便能准确重复研究过程。

对经济学家而言，一个重要议题是怎样记录和评估互联网上的交易行为。需要谨记保护隐私是个人权利，要解决数据所有权、数据托管和数据隐私的问题。[23] 应该完善数据供应的制度结构，以避免少数公司垄断数据。在多数情况下，数据并非是大范围开放的。另外，也有许多关于政府如何使用公民数据的问题。互联网数据也能用于经济决策。然而，银行可以实时监测客户的交易行为，客户的数据保护就难以保证了。麦克拉伦和珊布格解释了国家银行可以怎样通过网络搜索数据进行经济即时预报。[24]

互联网数据可以用于解决很多领域的人力资源问题，包括：即时预报（比起传统的数据搜集渠道，相关信息能更快获得），如麦克拉伦和珊布格[25]、阿斯吉塔斯和兹姆曼[26]、加黑艾尔-斯沃洛（Carrière-Swallow）和拉比（F. Labbé）[27]以及陈（T. Chen）等的研究[28]；预测（比如预测失业率、商品消费量、游客到访数和节日赛事赢家），如阿斯吉塔斯和兹姆曼[29]、富森（S. Vosen）和施密特（T. Schmidt）[30]、蔡（H. Choi）和范里安（H. Varian）[31]、阿桃拉（C. Artola）等[32]的研究；发现卫生与福利问题（抑郁、流感、经济危机时期的贫困），如金斯伯格（J. Ginsberg）等[33]、阿尔伯特·杨（A. C. Yang）等[34]、泰福特（N. Tefft）[35]、拉泽（D. Lazer）等[36]、阿斯吉塔斯和兹姆曼的研究[37]；记录不同生活情境中的匹配过程（例如寻找伴侣、找工作、购物），如阿斯吉塔斯和兹姆曼[38]、库恩（P. Kuhn）和曼苏尔（H. Mansour）[39]、库恩[40]、Kureková 等人[41]的研究；在传统数据存在缺陷的情况下对复杂系统进行评估（如发展中国家的集体谈判协议、跨国移民），如黑茨（G. J. Hitsch）等[42]、瑞普斯（U. D. Reips）和布发迪（L. E. Buffardi）[43]、比拉里（F. Billari）等[44]、比萨姆斯卡（J. Besamusca）和提登思（K. Tijdens）[45]，以及白楼（A. Bellou）[46]的研究。

三、利用互联网数据对失业进行建模

在多数情况下，宏观经济时间序列数据的发布都相当滞后，而且可能会遭到各种改动。同样，失业数据的发布也会滞后。因此，对实时掌握失业数据动态的需求日益强烈。[47]

欧盟委员会要求欧盟国家提供用于经济分析的大量数据。这些数据是从许多基于普查和抽样的大型调查中获得的。2005 年 8 月，欧盟委员会关于短期商业数据统计的规定中对此有明确要求，这与欧洲央行和欧洲统计局对欧洲货币联盟统计要求行动计划的发布相呼应，并且得到了欧盟成员国统计机构的支持。欧盟经济体短期经济分析所需的重要指标在创制和传播阶段要花费很长时间，欧盟现在之所以对数据统计提出要求，主要就是为了缩短这个时长。[48]

由于近期的经济金融危机，经济大幅下滑，失业这个宏观经济指标就成了大众和研究者都特别关注的对象。

经济大衰退期间，人们需要关于失业的短期数据，但却找不到。阿斯吉塔斯和兹姆曼于 2009 年开创性地指出，德国的月均失业率和特定的谷歌检索关键词高度相关。[49] 根据观察到的结构，他们预测了在即将来临的大衰退之复杂变幻的情况下，失业状况究竟如何。阿斯吉塔斯和兹姆曼采用时间序列数据的格兰杰因果检验方法，通过其他相关变量的波动，对德国的月均失

业率做出解释。他们还采用 2004 年 1 月至 2009 年 4 月未经季节性调整的数据建构了误差修正模型。他们尝试了不同的检索关键词，例如“unemployment rate”（失业率）、“unemployment office or agency”（失业救济办公室/机构）、“most popular search engines in Germany”（德国最受欢迎的搜索引擎）和“personnel consultant”（人事顾问）。

在另一项研究中，阿斯吉塔斯和兹姆曼改进了关键词，对更新后的模型进行了重新估计，来研究失业分析及预测的质量究竟如何，并把它和主要的竞争性劳动力市场指标做了比较。[50] 我们在这里简要概括一下这种方法，并对该研究策略的步骤和主要贡献做介绍。其核心回归方程是著名的误差修正模型（Y 代表失业率，X 是指标矢量）：

$$\Delta Y_t = \alpha + \beta y_{t-12} + \sum_{i=1}^{k} (\gamma_i \Delta X_{i,\,t} + \delta_i X_{i,\,t-12}) \quad (1)$$

其中，$\Delta Y_t = Y_t - Y_{t-12}$，$\Delta X_t = X_t - X_{t-12}$，$\Delta$ 为滞后 12 期的差分项，$t = 1, 2, \ldots, n$。

互联网活动指标或说检索关键词都是和“labour office”（劳动局）、“short-term work”（短期工）、“jobsearch”（找工作）相关的。相关技术细节可见阿斯吉塔斯和兹姆曼使用经济大衰退初期德国数据所做的研究。[51] 表 1 概括了基于最小二乘法（OLS）的估计结果，还包括了校正决定系数（R^2）、贝叶斯信息准则（BIC）、平均绝对误差（MAE）等评估手段。从前三行可以看到，互联网指标与实际失业率高度相关，其中同时包含三个指标的模型表现最佳（模型预测 2）。该模型中 R^2 等于 0.943，估计系数都呈现出统计显著性，正负号表明指标对失业率影响的方向，即搜索工作减少了失业，而搜索关于劳动局的信息以及通过短期工作获取资金支持，与失业率上升有关。这些发现很能说明问题。

除了上述 3 个互联网检索的关键词之外，还有两个传统的劳动力市场指标：Ifo-BB 和 DAX。其中，Ifo-BB 是由位于慕尼黑的 IFO 经济研究院根据公司个体数据推出的一个知名的传统劳动力市场指标，经常作为基准变量用于劳动力市场预测。DAX 是德国股票市场指数。阿斯吉塔斯和兹姆曼已经指出，滞后期为一年的 DAX 有同样好的预测功能，而且与 Ifo-BB 高度相关。这两个指标都能很好地反映德国失业率。关于 DAX，可参见表 1 第 7 行；关于 Ifo-BB，参见第 11 行。不过，它们的表现却远逊于第 2 行纯粹的互联网活动模型，这一模型使用了各类相关互联网数据。但是，互联网变量的预测质量在吸纳传统变量后会有所提升，这也是事实。根据 BIC 结果（见表 1），如果加上 Ifo-BB，第 2 行（只涉及所有即三个互联网变量）的 BIC 值可由 28.8 降至 11.4（第 9 行）；如果加上 DAX，则可降至 3.2（见第 6 行）。

表 1　　回归模型与领先一步预测结果

模型	劳动局	短期工	找工作	Ifo-BB	DAX	R^2 – a	BIC	MAE
预测 1	L*** + K +		L*** – K –			0.862	69.010	0.434
预测 2	L*** + K* +	L** + K*** +	L*** – K*** –			0.943	28.802	0.354
预测 3		L*** + K*** +	L*** – K –			0.923	38.986	0.420
预测 1+DAX	L*** + K +		L*** – K*** –		L*** – K*** –	0.950	21.589	0.263
预测 2+DAX	L*** + K +	L + K*** +	L*** – K*** –		L*** – K* –	0.969	3.178	0.297
预测 3+DAX		L*** + K** +	L*** – K*** –		L*** – K* –	0.955	16.177	0.429

续前表

模型	劳动局	短期工	找工作	Ifo-BB	DAX	R^2-a	BIC	MAE
DAX-预测					L*** − K*** −	0.887	53.216	0.314
预测 1+ifo-BB	L*** + K** +		L*** − K*** −	L*** − K*** −		0.950	21.645	0.333
预测 2+ifo-BB	L*** + K** +	L + K***+	L*** − K*** −	L*** − K** −		0.963	11.368	0.414
预测 3+ifo-BB		L*** + K +	L*** − K* −	L*** − K* −		0.938	32.593	0.550
ifo-BB-预测				L*** − K*** −		0.863	63.213	0.541

注：改编自 Askitas，N.，and K. F. Zimmermann. “Googlemetrie und Arbeitsmarkt”. *Wirtschaftsdienst*，2009，89（7）：495。数据来自 Arbeitsamt. de、IFO 经济研究院以及 Google Insights。Ifo-BB：慕尼黑 IFO 经济研究院的就业指标。DAX：德国股票市场指数。所用官方月均失业率数据未经季节性调整，但是在模型中已通过滞后 12 期对季节性进行了调整。要了解更多关于关键词的信息，可参见阿斯吉塔斯和兹姆曼的论文。公式（1）的所有标准回归模型涉及的数据都是 2005 年 1 月至 2009 年 5 月的月度数据。K 代表变化，L 指相关变量水平的 12 期滞后。+、−是估计系数的符号。* 代表统计显著性（* $P<0.05$，** $P<0.01$，*** $P<0.001$）。领先一步预测涉及的是 2008 年 3 月至 2009 年 6 月这个时段。R^2-a 是校正决定系数，BIC 是贝叶斯信息准则，MAE 指平均绝对误差。

从这一历史个案可以看出，互联网活跃数据确实蕴藏着有价值、有用且可用的信息。但是，我们需要在该项新技术的使用上积累更多经验，并且观察新数据在多大程度上可以替代传统信息来源。目前并不能想当然地认为我们能用互联网数据取代传统数据。

阿斯吉塔斯和兹姆曼用互联网活跃数据对失业进行建模的想法[52]也为其他国家的研究者所效仿。经验研究表明，在解释失业方面，与经济周期指标或传统时间序列模型相比，谷歌或其他互联网活跃数据能够补充额外的有用信息。类似的研究，有关于英国[53]、法国[54]、以色列[55]、意大利[56]、挪威[57]、土耳其[58]、巴西[59]的失业率，以及失业率水平，如西班牙[60]和乌克兰[61]，美国的失业救济申请[62]，关于谷歌和中国百度失业搜索指标的研究[63]。根据格兰杰因果检验，与失业相关的检索指标对于提高中国各种宏观经济指标的预测水平也有帮助。[64]

在谷歌活跃数据可用之前，艾特睿智（M. Ettredge）等人采用的是从 WordTracker “排名前 500 关键词报告”中提取的互联网搜索引擎关键词使用数据。[65]这一报告由金河联合有限公司（Rivergold Associates Ltd）每周发布一次。它涵盖了网络上最大的元搜索引擎（meta-search engines）。作者采用了 6 个最可能为找工作的人使用的表述，并以此预测美国的失业率，它们分别是：jobs（工作）、job listings（招聘启事）、namely job search（namely 找工作）、resume（简历）、employment（就业）和 monster. com（巨人招聘网站）。

以上研究多数使用了大量的谷歌检索数据。为减少数据维度，必须提取出一些主要成分。这些成分被作为解释变量用于像 ARMAX 这样的模型中。蔡和范里安在其研究中选择了两个指标：“welfare & unemployment”（福利与失业）和“jobs”（工作）。[66]他们发现，在美国，与失业、福利相关的检索可以提升对早期失业救济申请的预测质量。达木瑞（F. D’Amuri）和马库斯（J. Marcucci）在研究美国的情况时，只用了一个关键词：“jobs”（工作）。他们发现，谷歌指数（互联网工作搜寻指标）在预测美国失业率方面是最好的领先指标。[67]在研究德国的状况时，阿斯吉塔斯和兹姆曼用到了四组关键词，每组有一到八个词，中间用“或”运算符相连。[68]对于西班牙的情况，维森特等人则通过在 Google Trends 上查询“oferta de trabajo”（工作）和“oferta de empleo”（工作机会）加以了解。[69]

对于意大利的情况，纳卡拉拖（A. Naccarato）

等人分析了劳动力调查公布的官方失业率与Google Trends对“offerte di lavoro”（工作机会）的查询结果之间存在的协整关系。[70]在此前关于意大利的研究中，达木瑞和马库斯发现，“offerte di lavoro”是意大利人工作搜寻时最常用的关键词。[71]纳卡拉拖等人的研究表明，谷歌检索对于意大利失业率的即时预报很有用。[72]此前，弗兰塞斯库（D. A. Francesco）也同样使用了关键词“offerte di lavoro”，发现基于谷歌检索数据的模型能够完善对意大利失业率的样本外预测。[73]

而且，柏瑞拉（N. Barreira）等人分析了谷歌搜索在更多西南部国家的有效性，其结论是Google Trends数据有助于提升对意大利、法国和葡萄牙失业状况的分析，但西班牙却是例外。[74]他们使用的关键词和失业及失业救助有关。在研究意大利时，使用的关键词有：“disoccupazione”（失业）、“disoccupazioneordinaria”（失业救济金）和“INPS disoccupazione”（INPS失业救助，INPS是意大利国家社会保障局）。在研究法国时，使用的关键词有：“chomage”（失业）、“indemnites de chomage”（失业津贴）、“allocations chomage”（失业补助）和“allocations de chomage”（失业救济金）。在研究葡萄牙时，关键词用的是“desemprego”（失业）和“subsidiodesemprego”（失业补贴）。研究西班牙时，则是用到了“desempleo”（失业）、“subsidio de desempleo”（失业津贴）和“prestaciondesempleo”（失业条款）。

麦克拉伦和珊布格利用自回归模型，分析了英国官方失业率和一些检索词数据之间的关系，这些检索词包括：“unemployment”（失业）、“jobs”（工作）、“unemployed”（下岗）、“JSA”（失业救济金）、“Jobseeker's Allowance”（失业救济金）和“unemployment benefit”（失业福利）。作者证明，和既有调查相比，这些搜索数据包含有用的信息。JSA模型比只采用官方失业数据的基准模型能更好地解释失业问题。[75]

丰德尔（Y. Fondeur）和卡拉姆（F. Karamé）建构了经过卡尔曼滤波器和最大似然估计方法处理过的不可观察成分模型。通过这样的模型，可以复原不可观察成分，并估计未知参数。作者使用的变量是谷歌指数，以及15～24岁之间法国申请失业救济的人数。[76]

在转型国家，互联网应用有限，识字率也低，西方模型就难以适用。对于乌克兰来说，奥利克散德（B. Oleksandr）就没能证实互联网数据对解释失业率有用。[77]不过，一旦互联网在乌克兰经济生活中发挥更加重要的作用，这种状况就可能会发生变化。或者这也可能是由于没找到成功的研究策略。要知道，随着时间的推移，互联网数据结构的稳定性对发达国家而言都可能是有局限性的。那么，在转型国家和发展中国家遇到的挑战就更大了。但是，对于传统数据和模型来说，这些挑战同样存在。

帕夫利赛克（J. Pavlicek）和克里斯托法克（L. Kristoufek）分析了2004年1月至2013年12月维谢格拉德集团四国（Visegrad countries，即捷克共和国、匈牙利、波兰和斯洛伐克）月均失业率和与工作相关的查询之间的关系。[78]结果表明，谷歌搜索只在解释捷克和匈牙利的失业率方面有用。这可能是因为捷克和匈牙利有很多人移居境外，对在国外找工作感兴趣。波兰和斯洛伐克的情况究竟如何，还有待研究。

同时，对于巴西这个新兴经济体的情况，拉索（F. Lasso）和斯尼德斯and（S. Snijders）的研究发现，谷歌检索与失业之间高度相关，但季节性模型的影响更大。[79]他们使用的关键词是：“empregos”（工作）、“segurodesemprego”（失业保险）、“décimoterceirosalário”（第13个月工资）、“FGTS”（遣散费赔偿基金）、“INSS”（国家社会保障局）、“job vacancies index”（就业机会指数）、“unemployment and social benefits index”（失业和社会福利指数）。在研究土耳其的情况时，查德威克和桑谷尔使用的关键词是：“unemployment”（失业）、“unemployment insurance”（失业保险）、“job announcements”（招聘启事）、“looking for a job”（找工作）、“cv”（简历）、“career”（职业）。在贝叶斯模型平均的框架下，作者发现，谷歌检索数据只对土耳其非农业部门月均失业率的即时预报有效。其失业率官方数据来自家庭劳动力调查报告（Household Labor Survey），互联网数据则是通过Google Insights for Search搜集的。

四、结论

近年来，由于互联网数据的可用性，研究者开始使用这些数据来分析或预测宏观经济指标。这可能不仅仅由于互联网数据易得、丰富、经济、数字化，还可能因为互联网已经成为个人日常生活的一部分，能越来越多地反映现实行为趋势。

对失业情况变化的估计，既有研究大多依赖官方渠道，或者可能并不总是可靠的调查报告。而且，在发展中国家，主管机构常常出于各种原因而无法提供有价值的宏观经济指标评估，比如失业评估。多数关于失业即时预报的既有研究分析的是发达国家，如美国、英国、意大利、德国、芬兰或比利时。少数研究涉及了公共机构较弱的非西方国家，如维谢格拉德集团四国、乌克兰、土耳其和巴西。

本文分析了互联网数据在不同领域的应用，集中探讨的是它们在失业建模上的应用。本文提到的经验研究表明，互联网数据应用存在巨大的潜力，需要进一步挖掘。对多数国家而言，互联网数据能够改进失业模型和失业预测分析。不过，预测的准确度还要取决于一个国家的互联网普及率、互联网用户的年龄结构，以及所建构的互联网变量的稳定性。

参考文献

[1] Edelman, B. "Using Internet Data for Economic Research". *The Journal of Economic Perspectives*, 2012, 26 (2): 189 - 206.

[2] Askitas, N. "Social media: eine technologische und ökonomische Perspektive". In Rogge, C., and R. Karabasz (eds.). *Social Media im Unternehmen-Ruhm oder Ruin*. Wiesbaden: Springer Vieweg, 2014: 155 - 166.

[3] Arthur, W. B. "The Second Economy". *McKinsey Quarterly*, 2011 (4).

[4] [23] Askitas, N., and K. F. Zimmernann. *Detecting Mortgage Delinquencies*. IZA DP 5895, IZA, Bonn, 2011.

[5] [6] [9] [11] [29] [38] [49] [52] [68] Askitas, N., and K. F. Zimmermann. "Google Econometrics and Unemployment Forecasting". *Applied Economics Quarterly*, 2009, 55 (2): 107 - 120.

[7] [12] [17] [37] Askitas, N., and K. F. Zimmermann. "Health and Well-being in the Great Recession". *International Journal of Manpower*, 2015, 36 (1): 26 - 47.

[8] Constant, A., and K. F. Zimmermann. "Im Angesicht der Krise: US-Präsidentschaftswahlen in transnationaler Sicht". *DIWWochenbericht*, 2008, 44: 688 - 701.

[10] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and D. J. Watts. "Predicting Consumer Behavior with Web Search". *Proceedings of the National Academy of Sciences*, 2010, 107 (41): 17486 - 17490.

[13] [24] [25] [53] [75] McLaren, N., and R. Shanbhogue. "Using Internet Search Data as Economic Indicators". *Bank of England Quarterly Bulletin*, 2011 (2).

[14] [20] Zagheni, E., and I. Weber. "Demographic Research with Non-representative Internet Data". *International Journal of Manpower*, 2015, 36 (1): 13 - 25.

[15] European Union Internet Statistics. "Internet Usage in the European Union, 2016". http://www.internetworldstats.com/stats9.htm.

[16] [26] Askitas, N., and K. F. Zimmermann. "Nowcasting Business Cycles Using Toll Data". *Journal of Forecasting*, 2013, 32 (4): 299 - 306.

[18] [40] Kuhn, P. J. "The Internet as a Labor Market Matchmaker". *IZA World of Labor*, 2014, 18 (5): 1 - 10.

[19] [42] Hitsch, G. J., Hortaçsu, A., and D. Ariely. "Matching and Sorting in Online Dating". *The American Economic Review*, 2010, 100 (1): 130 - 163.

[21] [31] Choi, H., and H. Varian. "Predicting the Present with Google Trends". *Economic Record*, 2012, 88 (s1): 2 - 9.

[22] Askitas, N. "Google Search Activity Data and Breaking Trends". *IZA World of Labor*, 2015.

[27] Carrière-Swallow, Y., and F. Labbé. "Nowcasting with Google Trends in an Emerging Market". *Journal of Forecasting*, 2013, 32 (4): 289 - 298.

[28] Chen, T., So, E. P. K., Wu, L., and I. K. M. Yan. "The 2007 - 2008 US Recession: What Did the Real-Time Google Trends Data Tell the United States?". *Contemporary Economic Policy*, 2015, 33 (2): 395 - 403.

[30] Vosen, S., and T. Schmidt. "Forecasting Private Consumption: Survey-based Indicators vs. Google Trends". *Journal of Forecasting*, 2011, 30 (6): 565 - 578.

[32] Artola, C., Pinto, F., and P. de Pedraza. "Can Internet Searches Forecast Tourism Inflows?". *International Journal of Manpower*, 2015, 36 (1): 103 - 116.

[33] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and L. Brilliant. "Detecting Influenza Epidemics Using Search Engine Query Data". *Nature*, 2009, 457 (7232): 1012 - 1014.

[34] Yang, A. C., Huang, N. E., Peng, C. K., and S. J. Tsai. "Do Seasons have an Influence on the Incidence of Depression? The Use of an Internet Search Engine Query Data as a Proxy of Human Affect". *PloS one*, 2010, 5 (10): e13728.

[35] Tefft, N. "Insights on Unemployment, Unemployment Insurance, and Mental Health". *Journal of Health Economics*, 2011, 30 (2): 258 - 264.

[36] Lazer, D., Kennedy, R., King, G., and A. Vespignani. "The Parable of Google Flu: Ttraps in Big Data Analysis". *Science*, 2014, 343 (6176): 1203 - 1205.

[39] Kuhn, P., and H. Mansour. "Is Internet Job Search still Ineffective?". *The Economic Journal*, 2014, 124 (581): 1213 - 1233.

[41] Kureková, L. M., Beblavy, M., and A. E. Thum. "Using Internet Data to Analyse the Labour Market: a Methodological Enquiry". IZA Discussion Papers, 2014, No. 8555

[43] Reips, U. D., and L. E. Buffardi. "Studying Migrants with the Help of the Internet: Methods from Psychology". *Journal of Ethnic and Migration Studies*, 2012, 38 (9): 1405 - 1424.

[44] Billari, F., D'Amuri, F., and J. Marcucci. "Forecasting Births Using Google". Annual Meeting of the Population Association of America, PAA, New Orleans, LA, 2013.

[45] Besamusca, J., and K. Tijdens. "Comparing Collective Bargaining Agreements for Developing Countries". *International Journal of Manpower*, 2015, 36 (1): 86 - 102.

[46] Bellou, A. "The Impact of Internet Diffusion on Marriage Rates: Evidence from the Broadband Market". *Journal of Population Economics*, 2015, 28 (2): 265 - 297.

[47] [54] [76] Fondeur, Y., and F. Karamé. "Can Google Data Help Predict French Youth Unemployment?". *Economic Modelling*, 2013, 30: 117 - 125.

[48] [70] [72] Naccarato, A., Pierini, A., and S. Falorsi. "Using Google Trend Data to Predict the Italian Unemployment Rate (No. 0203)". Department of Economics-University Roma Tre, 2015.

[50] [51] Askitas, N., and K. F. Zimmermann. "Googlemetrie und Arbeitsmarkt". *Wirtschaftsdienst*, 2009, 89 (7): 489 - 496.

[55] Suhoy, T. *Query Indices and a 2008 Downturn: Israeli Data*. Bank of Israel, 2009.

[56] Naccarato, A., Pierini, A., and S. Falorsi. "Using Google Trend Data to Predict the Italian Unemployment Rate (No. 0203)". Department of Economics-University Roma Tre, 2015; D'Amuri, F. *Predicting Unemployment in Short Samples with Internet Job Search Query Data*. University Library of Munich, Germany, 2009.

[57] Anvik, C., and K. Gjelstad. "Just Google It. Forecasting Norwegian Unemployment Figures with Web Queries". Working Paper, 11, Center for Research in Economics and Management, Oslo, 2010.

[58] Chadwick, M. G., and G. Sengül. "Nowcasting the Unemployment Rate in Turkey: Let's Ask Google". *Central Bank Review*, 2015, 15 (3): 15.

[59] [79] Lasso, F., and S. Snijders. "The Power of Google Search Data; An Alternative Approach to the Measurement of Unemployment in Brazil". *Student Undergraduate Research E-journal*, 2016 (2).

[60] [69] Vicente, M. R., López-Menéndez, A. J., and R. Pérez. "Forecasting Unemployment with Internet

Search Data: Does it Help to Improve Predictions when Job Destruction Is Skyrocketing?". *Technological Forecasting and Social Change*, 2015, 92: 132 - 139.

[61] [77] Oleksandr, B. *Can Google's Search Engine be Used to Forecast Unemployment in Ukraine*. Doctoral dissertation, Kyiv School of Economics, 2010.

[62] Choi, H., and H. Varian. "Predicting Initial Claims for Unemployment Benefits". *Google Inc*, 2009: 1 - 5; Choi, H., and H. Varian. "Predicting the Present with Google Trends". *Economic Record*, 2012, 88 (s1): 2 - 9.

[63] [64] Su, Z. "Chinese Online Unemployment-related Searches and Macroeconomic Indicators". *Frontiers of Economics in China*, 2014, 9 (4): 573 - 605.

[65] Ettredge, M., Gerdes, J., and G. Karuga. "Using Web-based Search Data to Predict Macroeconomic Statistics". *Communications of the ACM*, 2005, 48 (11): 87 - 92.

[66] Choi, H., and H. Varian. "Predicting Initial Claims for Unemployment Benefits". *Google Inc*, 2009: 1 - 5.

[67] D'Amuri F., and J. Marcucci. "Google It! Forecasting the US Unemployment Rate with a Google Job Search Index". ISER Working Paper Series, No. 2009 - 32, 2009.

[71] D'Amuri, F. *Predicting Unemployment in Short Samples with Internet Job Search Query Data*. University Library of Munich, Germany, 2009; D'Amuri F., and J. Marcucci. "Google It! Forecasting the US Unemployment Rate with a Google Job Search Index". ISER Working Paper Series, No. 2009 - 32, 2009.

[73] Francesco, D. A. "Predicting Unemployment in Short Samples with Internet Job Search Query Data". *MPRA Paper*, 18403, 2009: 1 - 18.

[74] Barreira, N., Godinho, P., and P. Melo. "Nowcasting Unemployment Rate and New Car Sales in South-western Europe with Google Trends". *NETNOMICS: Economic Research and Electronic Networking*, 2013, 14 (3): 129 - 165.

[78] Pavlicek, J., and L. Kristoufek. "Nowcasting Unemployment Rates with Google Searches: Evidence from the Visegrad Group Countries". *PloS one*, 2015, 10 (5): e0127084.

Big Data and Unemployment Analysis

Mihaela Simionescu[1], Klaus F. Zimmermann[2]

(1. Institute for Economic Forecasting, the Romanian Academy, Bucharest;
Centre for Migration Studies, Prague Business School, Prague.
2. Princeton University, Princeton; UNU-MERIT & Maastricht University)

Abstract: Internet or "big" data are increasingly used in measuring the relevant activities of individual, households, firms and public agents in a timely way. The information set involves large number of observations and embraces flexible conceptual forms and experimental settings. Therefore, internet data are extremely useful to study a wide variety of human resource issues, including forecasting, nowcasting, detecting health issues and well-being, capturing the matching process in various parts of individual life, and measuring complex processes where traditional data have known deficits. This paper focuses on the analysis of unemployment by means of internet activity data, a literature starting with the seminal article of Askitas and Zimmermann. The article provides insights and a brief overview of the current state of research.

Key words: big data; unemployment; internet; Google; internet penetration rate

(责任编辑　武京闽)