失业率预测研究
——基于网络搜索数据及改进的逐步回归模型

●彭 赓 苏亚军 李 娜

摘要:无论是从社会管理还是从经济发展的角度来考虑,失业均已成为目前各国十分关注的重点问题之一,学者们也一直在通过各种方法来预测失业率。近年来,随着网络的发展和搜索引擎的普及应用,学者们发展出一种利用网络搜索数据来观察和研究经济及社会问题的方法。随着这一方法的有效性被证明之后,它也被引入到失业率预测的研究领域中。文章利用 Google 推荐的关键词搜索数据,采用改进的逐步回归方法分层建立了三个模型预测失业率,并进行因果关系检验及有效性检验。实验结果表明,三个模型的拟合优度分别达到 0.930、0.935、0.936,三期预测值的 MAPE 分别为 1.20%、0.89%、0.57%。文章认为,这种方法能有效的处理网络搜索数据并进行相关的社会问题研究和经济问题的预测。

关键词:逐步回归;失业率;失业初请人数;网络搜索数据;协整分析;预测

一、引言

本文以经济复苏时期美国失业率预测为例,从关键词库的构建、利用改进的逐步回归的方法对关键词进行筛选、合成综合搜索指数、构建模型进行预测等方面进行系统介绍,并对该方法的有效性及预测效果加以验证。

二、文献综述

1. 利用搜索数据进行失业率预测的相关研究。作为反映经济表现的指标之一,失业率一直受到非常广泛的关注。在失业率的预测方面,学者们采用各种方法用以提高预测的准确性。在 Ginsberg 等人利用 Google 搜索数据成功的预测美国流感疾病趋势以后,这一预测方法被迅速的引入到失业率预测中来。Askitas 和 Zimmermann 等(2009)建立了搜索数据与德国失业率之间的关联关系,并发现失业率发生变化时,网民对国家劳动局或失业保障机构、人事顾问、流行职位的搜索关注度会有所反应。D´Amuri 和 Marcucci 等(2009)利用 Google 搜索数据建立了工作搜索指数来预测美国的失业率,并发现在加入了搜索指数修正之后的模型的预测效果显著高于传统模型。Suhoy(2009),Choi 和 Varian(2009)将网络搜索数据加入到长期和短期的失业初请人数预测模型,发现模型的拟合度有较大的提高,并且在长期预测模型和短期预测模型中,过去 24 周的滚动预测值平均绝对误差分别降低 15.74% 和 12.90%。Wei Xu 和 Ziang Li 等(2012)利用网络搜索数据和神经网络方法构建美国失业率预测模型,发现这种模型比其他的预测模型的效果要更好。

2. 搜索数据关键词的选取。在利用网络搜索数据进行社会和经济研究方面,面对的都是海量的搜索数据和关键词,如何从中筛选出有预测价值的关键词是一个核心问题。对于这一问题,学者们处理方法各不相同。

第一种是采取技术取词法,即利用高性能、大规模的计算设备将一切可能的关键词都纳入到研究范围内,然后将相关统计模型编成程序运算选出核心关键词。例如 Ginsberg 等人利用 800 余台高速计算机在 2003 年~2008 年间 5 000 万个最为常用的搜索词中选择出 45 个与 CDC 发布的流感病人就诊量数据相关性最高的关键词,作为预测关键词的来源。

第二种是经验取词法,即由作者运用主观经验确定关键词。例如 Askitas 在网络搜索与失业率相关性时,认为与劳动局或失业机关机构、失业率、人事顾问和德国比较流行的几个职业搜索引擎四类关键词的搜索量将出现变化,因而以这四类关键词为核心合成搜索指标。

第三种是范围取词法,即先确定一个选词的范围,然后在范围内进行精选。例如 Konstantin 在研究网络搜索与美国个人消费增长率的相关性时,首先收集了 Google 提供的 27 个分类中的前十大搜索词,然后分别做出相关性分析后剔除与个人消费不相关的词,利用剩下 220 个与消费相关的词合成一个指标(Konstantin,2009)。Wei Xu 和 Tingting Zheng 等从 Google Trends 中与失业相关的分类中

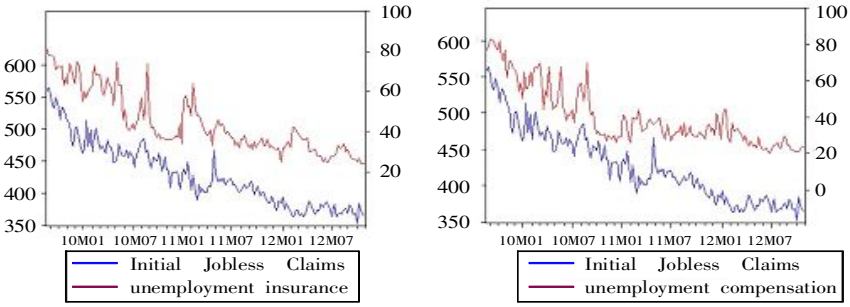

图 1 关键词搜索量变化与失业初请人数变化曲线图

表 1　ADF 平稳性检验

变量	ADF 检验	MacKinnon critical			ADF 检验结果
	t-statistic	1% level	5% level	10% level	
$LogY_t$	-2.061 752	-3.471 454	-2.879 494	-2.576 422	非平稳
$LogY_{t-1}$	-1.935 959	-3.471 454	-2.879 494	-2.576 422	非平稳
$LogQ_{2t}$	-1.898 558	-3.470 679	-2.879 155	-2.576 241	非平稳
$LogQ_{3t}$	-1.602 278	-3.471 192	-2.879 38	-2.576 361	非平稳
$LogQ_{4t}$	-1.411 442	-3.470 934	-2.879 267	-2.576 301	非平稳
$\Delta LogY_t$	-9.983 73	-3.471 454	-2.879 494	-2.576 422	平稳
$\Delta LogY_{t-1}$	-9.698003	-3.471 454	-2.879 494	-2.576 422	平稳
$\Delta LogQ_{2t}$	-17.442 08	-3.470 679	-2.879 155	-2.576 241	平稳
$\Delta LogQ_{3t}$	-10.718 54	-3.471 192	-2.879 38	-2.576 361	平稳
$\Delta LogQ_{4t}$	-12.634 75	-3.470 934	-2.879 267	-2.576 301	平稳

"Local/Jobs" 和 "Society/Social Services/Welfare & Unemployment" 类目中收集了 500 个左右的关键词作为原始关键词，从这 500 个关键词中找出相关系数大于 0.65 的 108 个关键词。然后利用神经网络方法，从这 108 个关键词中又筛选出少数几个关键词进行拟合

在现有取词方法中，技术选词法虽然精度较高，但是容易受到资源的限制而难以复制应用。而直接取词法与范围取词法虽然降低了工作量但是主观性较强，降低了学术研究的科学性(Ying Liu，Benfu Lv，2012)。

三、理论分析与预处理方法

随着网络的兴起与发展，人们越来越多的依靠网络来寻找所需要的各种信息。对于面临失业或是处于找工作的人而言，通过网络搜寻相关的工作信息无疑是一种非常便捷的渠道。当经济不景气，在岗的人出于对未来失业的担忧，会通过网络了解失业保障政策以及寻找工作机会。未在岗的人也会通过网络来了解经济形势及寻找工作机会，Google 把这些与失业相关的搜索关键词分成两类，分别是"工作"及"福利/失业"。虽然这些关键词的搜索数据量并不一定全是失业者或即将面临失业者的搜索产生的，但从统计上来说，这类关键词的搜索量与失业数据(这里的失业数据指的是初请失业人数数据，因为初请失业人数是美国失业率的非常好的先行指标-Choi and Varian(2009))，这两者之间存在很强的相关性。以"unemployment compensation"或"unemployment insurance"为例，其变化趋势和失业人数变化趋势存在很强的一致性。

1. 关键词选取方法。本文的关键词来源于 Google Trends(http://www.google.cn/trends/)，Google Trends 记录了从 2004 年以来某个关键词每一周被搜索的次数，并按照一定的算法将其标准化，并且，Google Trends 还根据搜索的关键词推荐与该关键词热门相关的一些关键词。本文采取的关键词选词方法是先选一个和失业相关的关键词作为初始关键词，由该关键词出发，Google Trends 会推荐出与其热门相关 10 个关键词，我们进一步搜索这 10 个关键词，Google Trends 又从这 10 个关键词出发又会推荐出其热门相关的大约 100 个关键词。如此往复，剔除重复的关键词后，这些关键词构成本文的基础关键词库。

本文将"unemployment"作为初始关键词，并且构成第一层的关键词库。第二层关键词库由第一层关键词库中的关键词及与这些关键词热门相关的 10 个关键词组成，去重后共 11 个，第三层关键词库由第二层关键词库的关键词及与这些关键词热门相关的关键词组成，去重后共 90 个，按照此方法，第四层关键词库共 403 个关键词。

2. 数据来源。本文采用的失业数据来源于美国劳工部网站 (http://www.ows.doleta.gov/unemploy/claims.asp)发布的经过季节调整后的每周初请失业人数，美国国家经济研究局(NBER)发布报告称从 2007 年 12 月份美国开始进入衰退期，2009 年 6 月份衰退结束，进入恢复期。本文重点研究经济恢复时期的美国失业率。所以，本文采用的数据跨度为 2009 年 6 月至 2012 年 10 月期间的 175 周，将前 172 周的数据作为训练集进入模型用于参数估计，后 3 周的数据作为测试集用于评估预测效果。

3. 搜索指数合成的方法。

(1)逐步回归法。在线性回归模型中，通常会碰到两个问题：一是如何从众多的自变量中挑选出对因变量有显著影响的解释变量。二是如何消除自变量之间存在的多重共线性对回归方程的影响。逐步回归分析方法被认为是解决这两个问题的有效方法之一。它的核心思想是在考虑的全部自变量中按其对因变量的贡献程度大小，由大到小地逐个引入回归方程中，如果发现先前被引入的自变量在其后由于某些自变量的引入而失去其重要性，可以从回归方程中随时予以剔除。直到既无不显著变量从方程中剔除，又无显著变量需要引入回归方程为止。其主要步骤如下：

Step1：对所有的自变量和应变量进行标准化处理；

Step2：计算自变量和因变量之间的皮尔逊相关系数，并找出相关系数最大的因变量，并根据偏 F 检验来判断该因变量是否应该被引入模型中；

表 2　回归结果及检验

模型 / 系数	模型 1		模型 2		模型 3	
	系数值	P 值	系数值	P 值	系数值	P 值
β_0	0.336 585	0.000 0	0.452 783	0.000 0	0.472 746	0.000 0
β_1	0.795 300	0.000 0	0.690 524	0.000 0	0.676 872	0.000 0
β_2	0.078 661	0.000 1	0.114 292	0.000 0	0.111 178	0.000 0
Adjusted R-squared	0.929 592		0.934 994		0.935 56	
Log likelihood	483.437 1		490.022 7		490.743 8	
AIC	-5.823480		-5.903 306		-5.912 046	
SC	-5.767 008		-5.846 834		-5.855 574	
DW	2.361 125		2.314 068		2.283 341	
残差平稳性检验						
ADF 值	-15.420 57		-15.051 67		-14.860 64	
1% level	-3.470 427		-3.470 427		-3.470 427	
检验结论	平稳		平稳		平稳	
结论	协整		协整		协整	

表 3 Granger 因果关系检验

模型	原假设	观察值	F 统计量	P 值
模型一	$LogQ_{2t}$ 不能 Granger 引起 $LogY_t$ $LogY_t$ 不能 Granger 引起 $LogQ_{2t}$	164	4.891 47 12.749 4	0.028 4 0.000 5
模型二	$LogQ_{3t}$ 不能 Granger 引起 $LogY_t$ $LogY_t$ 不能 Granger 引起 $LogQ_{3t}$	164	13.726 6 18.157 3	0.000 3 3.E-05
模型三	$LogQ_{4t}$ 不能 Granger 引起 $LogY_t$ $LogY_t$ 不能 Granger 引起 $LogQ_{4t}$	164	16.505 1 13.967 2	8.E-05 0.000 3

表 4 预测结果比较

	10/20/2012	10/27/2012	11/03/2012		
真实值(K) 预测值(K)	372	367	363	MAPE	RMSE
模型一	379.1	369.6	366.6	1.20%	4.82
模型二	380.0	368.7	362.7	0.89%	4.73
模型三	376.4	367.4	364.5	0.57%	2.69

Step3：在逐步引入新的因变量的同时，利用偏 F 检验删除之前进入模型但其对因变量的贡献降低的自变量；

Step4：重复 Step2 和 Step3 的过程，直到无显著变量需要引入回归方程为止。

(2)改进的逐步回归法。在利用网络搜索数据进行多元回归分析中，也需要从大量的关键词中筛选出对因变量有显著影响的自变量。不同的是，网络搜索的关键词和真实经济量变化之间存在先行或滞后的关系。由于要利用网络搜索数据进行预测，而因此必须找到那些搜索趋势变化领先于因变量变化的关键词(刘颖等，2011)。本文将这类关键词称为先行关键词。在运用逐步回归法筛选自变量之前，需要先利用时差相关分析法确定关键词的领先阶数，然后再利用逐步回归的思想合成综合搜索指数，本文将这一系列的处理过程称为改进的逐步回归分析法，其主要步骤如下：

Step1：利用时差相关分析法分析关键词的领先阶数。时差相关分析法是利用时差相关系数来验证经济时间序列先行或滞后关系的一种方法，其公式如下：

$$r_l=\frac{\sum_{t=1}^{n}(x_{t+l}-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{t=n}(x_{t+l}-\bar{x})^2\sum_{t=1}^{n}(y_t-\bar{y})^2}},(l=0,\pm1,\pm2\cdots)$$

上式中，r_l：时差为 l 的相关系数，y：因变量，$\bar{y}$：因变量均值，x：自变量，$\bar{x}$：自变量均值。l 为 x 的领先阶数。相关系数最大的时差阶数为领先阶数。此时的时差相关系数即为两者之间的相关系数；

Step2：确定了领先阶数后，将关键词按照领先阶数进行时差调整后与基准指标进行回归。将拟合度最大的关键词进入综合搜索指数。并对该综合搜索指数进行显著性检验；

Step3：对其他的关键词加入综合搜索指数之后，与基准指标进行再次回归，将对拟合度提升最大的关键词加入合成指数，形成新的综合搜索指数，并进行显著性检验；

Step4：对进入综合搜索指数的其他关键词再次进行显著性检验，如果不通过，则删除该关键词数据；

Step5：重复 Step3 和 Step4，直至拟合度不再提高时停止，我们将此时的综合搜索指数记为 Q_{it}(i=1,2,3,4 分别代表第一、二、三、四层关键词经过上述操作步骤后最后合成的综合指数)。这样就能持续地将能够显著提高搜索综合指数拟合度的关键词选取出来。

四、实证分析

1. 模型建立。本文将失业初请人数 Y_t 作为被解释变量，以提前一期的失业初请人数 Y_{t-1} 作为解释变量一，以搜索数据综合指数 Q_{it}(i=2,3,4)作为解释变量二分别建立三个模型来验证本文提出的关键词选取方法及搜索数据预处理方法的有效性（由于单个关键词的选取带有很大的随机性，所以不建立 i=1 时的模型)。为增进平稳性，降低异常数据的影响，本文分别对以上变量取对数，分别表示为 $LogY_t$，$LogY_{t-1}$，$LogQ_{it}$。

在建立模型之前，需要对各变量进行平稳性检验，本文采用 ADF 检验法对以上变量进行平稳性检验，检验结果如下：

从 ADF 检验结果来看，原变量序列均为非平稳序列。而一阶差分后的变量序列均为平稳序列。所以，以上变量均为一阶单整序列。

根据前面建立的理论框架，按照第二、三、四层的关键词综合指数，我们建立了如下模型。

$$LogY_t=\beta_0+\beta_1LogY_{t-1}+\beta_2LogQ_{2t}+u_t \quad (1)$$

$$LogY_t=\beta_0+\beta_1LogY_{t-1}+\beta_2LogQ_{3t}+u_t \quad (2)$$

$$LogY_t=\beta_0+\beta_1LogY_{t-1}+\beta_2LogQ_{4t}+u_t \quad (3)$$

模型回归结果及检验如表 2 所示。

在上述的三个模型中，各解释变量前的系数在 1%的水平上均显著不为零，说明本文建立的模型是合理的。搜索数据综合指数 $LogQ_{it}$ 的系数 β_2 显著为正，说明失业初请人数和失业相关的搜索关键词指数之间有显著的正相关关系。

进一步的，本文对构建的三个模型的残差进行平稳性检验，发现残差序列在 1%的水平上均具有平稳性。因此，解释变量和被解释变量之间存在一阶协整关系。

2. 因果关系检验。Granger 因果关系检验能检验解释变量的前期变化是否能有效的解释被解释变量的变化。因此，Granger 检验可以作为考察模型的预测能力的一个指标。本文对变量 $LogY_t$ 和 $LogQ_{it}$ 进行了一阶 Granger 因果关系检验。检验结果如表 3。

由表 3 可知，在三个模型中，解释变量　均可以显著地 Granger 引起被解释变量 $LogQ_{it}$，表明搜索综合指数确实能够对失业初请人数具有良好的预测效果。

3. 模型预测效果比较。为进一步考察模型的预测能力，本文运用以上 3 个模型，分别预测了 2012 年 10 月份后 3 周的失业初请人数，并以平均绝对百分误差 MAPE 和均方根误差 RMSE 作为衡量预测能力的指标。设 n 为预测值的个数，实际值为 y_i，预测值为 $\hat{y}_i$，则 $MAPE=\sum_{i=1}^{n}\frac{|y_i-\hat{y}_i|}{y_i}/n$ $RMSE=\sqrt{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2/n}$ 预测结果如表 4 所示。

从表 4 可知,三个模型均取得很好的预测结果。这说明,本文采取的基于 Google 推荐的热门相关关键词并利用改进的逐步回归法合成综合搜索指数的方法进行建模和预测是非常有效的,并且随着关键词层级的增加(即关键词库中关键词数的增加),模型的预测效果越好。

五、结论及展望

本文采用了一种新的网络搜索数据的处理方法进行美国失业率预测。在这种处理方法中,本文根据 Google 推荐的热门相关关键词构成关键词库,进而利用改进的逐步回归的方法合成搜索综合指数。从关键词的选择上来看,相较于技术选词法、经验选词法及范围选词法,本文中的选词方法更为简单和客观。从合成搜索综合指数的方法上来看,这种合成方法使得解释变量和被解释变量之间的皮尔逊相关系数非常高,平稳性也高。从实验结果来看,根据这种处理方法建立的模型和预测的效果都非常好,三个模型的预测平均绝对百分误差 MAPE 分别为 1.20%、0.89%和0.57%,均方根误差 RMSE 分别为 4.82、4.73 和 2.69。综上,这种对网络搜索数据进行处理和预测研究的方法能够显著有效地提高失业率预测的准确性。

目前,基于网络搜索数据进行经济和社会行为预测的研究都是关于应用方面的研究。对于网络搜索数据和经济与社会行为方面的内在机理进行探讨的文章不多,尚未形成系统的理论框架。在利用网络搜索数据进行失业率预测方面,处于经济的不同时期,人们进行搜索的行为模式是否会发生变化?网络搜索数据与传统的市场数据结合进行预测是否能取得更好的预测效果?此外,在所有采用搜索数据进行研究的相关文章中,搜索数据大部分都来源于 Google,而在一些发展中国家,Google 搜索引擎的市场份额较小,如何利用当地主流搜索引擎如百度的搜索数据进行经济和社会问题研究?以及基于这些数据进行的研究是否可以和基于 Google 的搜索数据进行的研究一样有效?这些问题,都是我们下一步的研究的内容与重点。

参考文献:

1. Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant, Detecting influenza epidemics using search engine query data, Nature,2009,(457):1012-1014.

2.N Askitas, and KF Zimmermann, Google Econometrics and Unemployment Forecasting. Applied Economics Quarterly,2009,55(2):107-120.

3. H Choi, H Varian, Predicting the Present with Google Trends, Technical Report, Google Inc, 2009.

4. 刘颖,吕本富,彭赓.互联网搜索数据预处理方法及其在股市分析中的应用. 情报学报,2011,10(10):1028-1036.

基金项目:国家自然科学基金项目“基于网络搜索数据的电子商务交易量预测研究——以 3C 产品为例”(项目号 71172199);国家自然科学基金项目“情绪因素对网络搜索预测的影响: 以旅游市场为例”(项目号:71202115);国家自然科学基金项目“基于网络行为的消费者信心指数构建及应用研究”(项目号:71203218);中国科学院研究生院院长基金项目“基于‘结构时序 + 搜索’混合模型的电子商务预测”(项目号:Y15101QY00)。

作者简介:彭赓,中国科学院大学管理学博士,中国科学院大学管理学院副教授;苏亚军,中国科学院大学管理学院硕士生;李娜,中国科学院大学管理学院硕士生。

收稿日期:2013-10-20。

(上接第 36 页)

受到严重影响。

熟知产品退市原因,识别企业弱势产品,科学、审慎地实施产品退市决策,是每个中国企业的必修课。然而相比国外,中国产品退市管理做得依然不够精细。以惠普为例,惠普的产品退市精细化、标准化程度非常之高,它规定了产品退市固定的流程,形成了很完整的体系,且惠普的产品经理要经过正规而严格的产品退市培训。此外惠普还有专门的退市团队,由销售经理、采购经理、企划经理等组成。在本文所研究的中国企业产品退市案例中,缺乏惠普这样的严谨产品退市体系,这说明中国企业在产品退市管理方面仍然需要向优秀企业借鉴经验,不断地提高自身的管理水平。

参考文献:

1. Linda Gorchels. The Product Manager's Handbook. 北京:中国财政经济出版社,2007.

2. Merle Crawford,Anthony Di Benedetto. New Products Management.北京:中国人民大学出版社,2006.

3. 桑赓陶. 一个被营销管理学术界忽视的重要研究领域:产品退市管理.市场营销导刊,2007,(6).

4. 李季,江明华. 构建正规化的产品退市管理体系的重要性及其实施要点.现代管理科学,2010,(6).

5. 张廷权.关于市场驱动的产品退市机制和策略的探讨.广东通信技术,2012,(8).

6. 周静,李季,江明华. 产品退市决策研究:基于 Cox 生存模型的实证分析.管理科学学报,2013,9(2):1-12.

7. 李志成.HTC、魅族巨人转身与单骑突围.商界(评论),2010,(9).

基金项目:教育部人文社会科学基金项目“构建产品退市决策的客观标准——基于 Cox 模型的实证研究”(项目号:09YJC630241);国家自然科学基金“谁来推荐更有效?顾客的社会网络对口碑推荐效果的影响机制研究”(项目号:71102127)。

作者简介:李季,中央财经大学商学院副教授,北京大学管理学博士;佟晓迪,中央财经大学商学院企业管理专业硕士生;朱彦奇,中央财经大学商学院市场营销专业本科生。

收稿日期:2013-10-08。