

基于网络搜索数据的 消费者信心指数预测研究^{*}

——以台湾地区为例

刘伟江 李映桥

提要:大数据时代网络中所蕴含的数据可以真实、及时地反映当前经济变化趋势。本文根据消费者行为理论构建消费者信心指数与网络搜索行为关系的概念框架,通过搜集与消费者信心指数构成相关的台湾地区关键词搜索量,运用主成分分析方法合成搜索指数并对该地区消费者信心指数进行预测。实证结果表明:由谷歌趋势中的关键词搜索量拟合出来的搜索指数与消费者信心指数之间存在长期稳定协整关系,模型拟合度高且时效性强,预测数据比官方公布的数据提前一个月左右。与传统的预测方法相比,模型具备转折点预测的能力。

关键词: 网络搜索数据 谷歌趋势 消费者信心指数 搜索量

作者刘伟江,吉林大学数量经济研究中心教授、博士生导师;李映桥,吉林大学数量经济研究中心博士研究生。(长春 130012)

一、引言

消费者信心指数(CCI)是一个国家或地区的消费者对当前经济状况满意程度和未来经济走向预期的综合指数,它是预测未来经济走势与消费者消费趋向的一个先行指标,该指标为国家政策制定、经济预测及其它领域应用提供了重要的参考依据。

台湾地区的消费者信心指数是由中央大学台湾经济发展研究中心主办,辅仁大学统计资讯系针对台湾的经济及社会等方面所涉及的经济形势、就业、物价水

平、生活状况、购房和投资等六类问题在每月第三周进行问卷调查,^①每月的最后一周公布计算结果。此结果由于调查和计算都需要时间,因此官方公布的数据会产生一定的时滞,而网络中网民的行为是即时的,合理运用网络搜索数据在某种程度上可以缩短这种时滞。同时网络的普及使网络搜索数据能反映大多数民众的意愿,以台湾网路资讯中心(TWNIC)每年发布的数据为例,2014年6月的《2014年台湾宽频网路使用调查》^②数据显示,台湾地区12岁以上的人群中有2556万人曾使用过网路,占全部人数的77.66%。如此众多的网民产生了大量的网络搜索数据,这些数据可以迅

* 祝蕾为本文提供了资源共享和数据处理等方面的关键帮助,特在此对她致以真诚的谢意。

① 中央大学台湾经济发展研究中心:《消费者信心指数调查报告》, <http://rcted.ncu.edu.tw/intro.phtml#5>。

② 台湾网络信息中心:《2014年台湾宽频网路使用调查》, <http://www.twonic.net.tw/download/200307/20140820e.pdf>。

速地反映人们的关注点和兴趣点。

本文从台湾地区民众主要使用的 Yahoo! 奇摩和谷歌趋势(Google Trends)中选取网络搜索关键词,研究网络搜索关键词与消费者信心指数之间的关联性,构建模型对消费者信心指数进行预测,该方法可以使预测结果早于调查结果并具有转折点预测的能力。文章的余下安排如下:第二部分对网络搜索数据和消费者信心指数方面的文章进行了概括与总结;第三部分建立了网络搜索数据与消费者网络搜索行为之间关系的概念框架;第四部分为数据来源及数据处理;第五部分构建了回归模型并对消费者信心指数进行预测和分析;第六部分为文章的结论。

二、文献综述

在国外的研究中,利用 Google Trends 所提供的关键词搜索量来预测经济指标虽未形成一套完整的研究体系,但网络搜索技术在研究失业率、预测流感发病率以及房地产销售等领域已有广泛的应用。

Eysenbach, Polgreen 根据人们在网络上对与流行性感冒相关的关键词搜索量来合成搜索指数,该指数与流感发病率之间存在长期稳定关系,进而利用该合成指数预测流感发病情况及流感的死亡率。^① Hyun-yong Choi 的研究发现美国劳动部公布的失业率与谷歌趋势上与失业密切相关的互联网搜索关键词的搜索量存在着非常显著的相关性,从而把网络搜索关键词当作变量引入模型中,进而提高了美国失业率预测模型的精度。^② Choi, Varian 对美国汽车及零配件、零售业、旅行行业和房地产等四个行业的销售量与网络搜索数据之间的关系进行了深入的研究,该研究在季节性自回归模型和固定效应模型基础上,将关键词网络搜索量作为一个新的影响因素加入到传统的自回归模型中并建立新的销售量预测模型,实证结果表明:四个行业的预测精确度都有了一定的提高,其中对汽车及零配件行业的销售量预测精度提高了 18%,房地产销量预测精度提高了 12%。^③ 杨欣、吕本富等以“7-23 甬温动车特大交通事故”为研究对象,从投资者视角构建

了重大突发事件对股价波动影响的理论框架,借助网络搜索数据形成市场搜索指数和事件搜索指数,建立移动自回归(ARMA)模型,通过实证分析探讨了突发事件对中国股市动车概念版的影响强度,实证结果表明网络搜索数据能够及时而又准确地反映该动车特大交通事故对相应板块股市波动的影响。^④ 张崇等研究通过获取的搜索数据构建模型来预测全国 CPI,模型拟合效果为 0.978,预测提前期为一个个月。^⑤ 这些研究在实证研究领域都有各自的突破,说明了网络搜索同社会经济行为之间存在相关性。然而,对于如何完整地收集关键词学术界还没有形成一套成熟的方法,多数研究在搜索指数的合成方面只是对所有关键词数据进行简单的算术平均,对于不同关键词应占据的权重有待进一步探究;在 Google Trends 的应用方面,使用中文作为搜索关键词进行预测其效用与使用英文关键词的效果是否有差异,目前仍缺乏相关研究证明。

三、消费者信心指数与网络搜索数据关系的概念框架

台湾地区 CCI 的数值是通过对其民众的问卷调查直接得出,因此其反映了台湾民众对经济满意程度及预期的最直观感受,这一感受在互联网中体现为在网络搜索、浏览等指标的变化方面。

根据对消费者信心的影响因素构成,把消费者在互联网搜寻的信息分为债务负担、能源成本、企业状况、耐用品需求、信贷需求和保险需求等六部分。它们对消费者信心的影响是:上升的债务负担会降低消费者的购买力与购买信心;上涨的燃气价格和燃油价格对家庭造成了压力;企业状况的变化会影响消费者的支出行为;耐用品需求的增加显示了消费者购买力的上升;信贷需求的增长则表示消费者对自己偿还能力的自信;保险需求的增加反映了消费者对潜在风险的规避和未来经济形势的不看好。这几方面在网民的搜索行为中都可以找到相对应的关键词,影响因素对民众的作用效果体现在相应关键词搜索量的改变,随着网络使用率的提高,这种作用效果在互联网世界是即时发生的。因此选择出合适的关键词进行处理并构建

① G Eysenbach, Infodemiology: Tracking Flu-related Searches on the Web for Syndromic Surveillance, *AMIA Annual Symposium Proceedings*, 2006, pp. 244-248. Philip M. Polgreen, Yiling Chen, David M. Pennock, Forrest D. Nelson, Using Internet Searches for Influenza Surveillance, *Clinical Infectious Diseases*, 2008, 47(11), pp. 1443-1448.

② Choi, Hyunyoung, and Hal Varian, Predicting initial claims for unemployment benefits, July 22, 2009, *working paper*, Google Inc.

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

④ 杨欣、吕本富等:《基于网络搜索数据的突发事件对股票市场影响分析》,《数学的实践与认识》2013 年第 23 期。

⑤ 张崇、吕本富等:《网络搜索数据与 CPI 的相关性研究》,《管理科学学报》2012 年第 7 期。

模型,就可以对 CCI 的走势与变动进行预测判断。

根据消费者行为理论,将消费者决策过程分为信息需求产生、信息搜寻和行为决策这三个部分。这样的决策过程需要大量的信息支撑,而互联网的日益发展与渗入使网络成为了人们获取信息的主要途径,资讯搜索代表民众的兴趣与关注,网页浏览反映了信息的传递,博客论坛反映了人们各自的看法与交流。本节从影响消费者信心的因素出发,分析 CCI 与网络搜索数据之间的关系,遵循如下思想构建概念框架:作为市场微观主体,消费者消费信心的变化会影响其消费观念和消费决策,而消费者的行为在经济市场与互联网世界都有所体现。在经济市场上的表现体现在其对商品的购买与投资,在互联网中的表现则体现为网络搜索及相关链接点击量的变化,那些对 CCI 变动起着主导作用的影响因素在人们的日常网络搜索中都可以找到相对应的关键词,影响因素对民众的作用效果体现在相应关键词搜索量的改变,但由于网络搜索关键词所表达的含义是不同的,即使是同样的关键词,它们所体现的消费者行为也可能是不同的,例如对电脑品牌的搜索,既可以解释为消费者希望通过网络帮助他们购买电脑,也可以解释为消费者只是在寻找该品牌电脑的在线升级。因此,就需要采用科学的方法对关键词进行筛选,选择指向单一、没有歧义的关键词,建立适当的模型,这样才能正确地 CCI 的走势与变动情况进行分析预测。基于上述分析,本文构建的 CCI 与网络搜索行为关系的概念框架如图 1 所示。

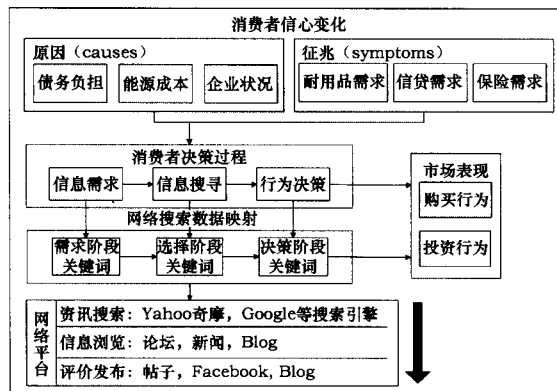


图 1 概念框架:CCI 与网络搜索行为的关系

四、数据来源及处理

(一)数据来源

本文所采用的消费者信心指数(CCI)数据来自于中央大学台湾经济发展研究中心按月发布的《消费者信心指数调查报告》(Taiwan Consumer Confidence Index(CCI) Report),本研究抽取了 2009 年 11 月至 2014 年 02 月共 52 个月的 CCI 同比数据。网络搜索关键词数据来自于谷歌趋势^①,这项服务由谷歌实验室在 2006 年 5 月推出,同年 7 月谷歌趋势中文版正式上线。利用这项服务可以获得基于谷歌网页搜索的某一给定关键词在台湾地区从 2009 年 11 月至当前周的周搜索量。需要注意的是,该搜索量的数值并不是绝对搜索量,而是测算用户在谷歌上对该主题的关注度,即“在选定时间、地区范围内您所选定的关键词在谷歌网页被搜索的次数占这段时间该地区谷歌搜索总量的百分比”^②。同时在谷歌趋势上还可以获得与选定关键词相关的关键词的热门推荐。

(二)关键词的选取与处理

1. 关键词的选取

首先,根据辅仁大学调查台湾消费者信心指数时采用的经济形势、就业、物价水平、生活状况、购房和投资等六项分类来选取相关的初始关键词。在选取过程中,鉴于台湾网民主要使用 Yahoo 奇摩与谷歌这两种搜索引擎进行日常搜索,本文通过“Yahoo! 奇摩知识”中的“知识分类”中收集这六大类下的精选关键词,如“折扣”、“汽车险”、“劳工局”、“景气”、“本益比”和“预售屋”等作为初始关键词;

然后,利用 Yahoo 奇摩以及 Google Trends 的相关热门搜索词推荐功能,依次输入每个初始关键词,选取并记录得到的推荐相关词,例如与“房价”相关的热门推荐词有“台北房价”、“查询房价”、“实价登录”等,在这里还可以继续深度搜索与推荐词相关的热门推荐词,例如与“实价登录”相关的热门推荐词有“实价查询”、“内政部”、“不动产”、“永庆”等,这种方法可以使我们尽可能全面地覆盖到民众对每个类别的兴趣与关注点。输入关键词后即可下载该关键词在选定区域(台湾)选定时间范围(2009 年 11 月至 2014 年 02 月)的周搜索量,数据以 CSV 文件形式保存。筛选搜索量不足的关键词后,从 Google Trends 上得到的关键词共 544 个。

2. 关键词预处理

首先,由于下载的关键词搜索量以周为单位呈现,而 CCI 为月度指标,二者频率不同,因此需要将周数据

① <http://www.google.com/trends/explore>.

② 译自 https://support.google.com/trends/answer/4355213?hl=zh-Hans&ref_topic=4365599.

合并为月数据;其次,为消除数据的短期波动,采取三期移动平均法对搜索数据进行平滑,突出其长期变化趋势。这时每个搜索关键词会损失两期(2009年11月与2014年02月)搜索量;然后将平滑后的数据转为同比数据从而与CCI指数保持一致。

数据进行如上处理后,使用stata统计软件,计算每个关键词搜索量(包括其 n 阶提前期)与CCI的相关系数,该系数可以反映各关键词搜索量与CCI之间相关关系的密切程度,系数绝对值越趋近于1表明该关键词与CCI间线性关系越密切。挑选出相关系数大于0.7的关键词共89个(其中相关系数大于0.8的搜索指数26个,0.7到0.8之间的关键词63个)。同时记录在该相关系数下的搜索指数滞后阶数。因为本文目的在于通过网络关键词搜索量预测CCI,故而在计算每个关键词搜索量与CCI的相关系数时只选取 $t-n$ 期 $n=0\sim 12$ 的关键词,不考虑 $t+n$ 期的情况。假如与第 t 期CCI相关系数最高的搜索指数为第 $t-12$ 期,则表明该搜索指数变化对CCI的传导时滞期为一年左右。在确认每个搜索关键词的时滞期后将所有搜索关键词按照其时滞期调整到与CCI同期影响,例如:“银行利率”对CCI存在12期时滞,则需要将2009年12月的“银行利率”搜索指数对应2010年12月的CCI,其余时点按此方式依次对应。

3. 搜索指数的合成

因为对时间序列的分析是通过建立以因果关系为基础的结构模型进行的,该分析有一个隐含的假设,即这些序列是平稳的。^①因此我们需要对CCI及各关键词搜索量进行平稳性(单位根)检验,使用ADF检验(Augmented Dickey-Fuller Test),原假设为序列至少存在一个单位根(不平稳),通过stata软件的相关检验命令得到CCI原序列为非平稳序列,但经过一阶差分后,检验结果在1%的显著水平下拒绝原假设,即时间序列不存在单位根,是平稳的。说明原序列为一阶单整序列。在此基础上我们依次检验所有关键词,排除自身平稳以及二阶及以上平稳的时间序列,保留所有一阶单整序列共62个。

由于在上述步骤中已经筛选了大量关键词,使得现存的关键词并不完全覆盖六项分类中的每一项,因此不应继续单纯地按照初期收集关键词时所依据的这六项分类来划分现有的62个搜索关键词;为了在保留所有强相关关键词的同时减少模型中的变量个数,根据现有关键词在社会经济生活中的属性重新将其划分

为五大类,即“就业”、“房地产”、“宏观经济”、“家庭经济”及“汽车”。而后采用主成分分析法对每一类别进行关键词搜索指数的合成,以期在减少数据集维数的同时,保持数据集中对方差贡献最大的主成分,从而保留住数据的主要信息。按照经主成分分析后给出的每类下各搜索指数占有的权重,加权得出每类所对应的解释变量。表1列出了关键词的五大类别和部分关键词的领先阶数及相关系数。

表1 五类搜索指数与部分关键词

就业类 指数(X_1)	房地产类 指数(X_2)	宏观经济类 指数(X_3)	家庭经济类 指数(X_4)	汽车类 指数(X_5)
职训中心 L12,0.736	优惠房贷 L6,0.704	银行贷款利率 L12,0.719	购物中心 L7,-0.748	Mercedes L9,0.725
职训局 L12,0.795	房仲网 L4,0.785	信贷 L12,-0.732	购物台东森 L12,0.772	Citroen L8,0.785
职训 L12,0.732	房市 L7,0.815	汇丰信用卡 L12,0.703	奇摩购物 L12,-0.743	Honda L9,0.821
劳工局 L12,0.759	房地产 L12,0.811	贷款 L4,0.853	购买 L12,0.702	Cadillac L10,0.779
就业e L12,0.784	中信房屋 L11,0.741	银行汇率 L10,0.845	acer笔电 L12,0.830	Peugeot L10,0.719
就业服务 L12,0.818	房屋 L11,0.789	富邦信用卡 L9,0.808	计算机 L11,0.736	Suzuki L11,0.767

五、模型建立及分析

(一)模型的建立

本文选择CCI作为被解释变量(Y),就业指数(X_1)、房地产指数(X_2)、宏观经济指数(X_3)、家庭经济指数(X_4)、汽车指数(X_5)作为解释变量,由于前面已对被解释变量进行过平稳性(单位根)检验,并知其为一阶单整序列。这里对各解释变量进行同样的检验,检验结果如表2所示,原序列均为非平稳序列,但经过一阶差分后,检验结果在1%的显著水平下拒绝原假设,即时间序列平稳的。说明它们均为一阶单整序列。

表2 变量平稳性检验

变量	ADF 检验 t-Stat	Mac Kinnon 临界值			ADF 检验结果
		1%	5%	10%	
Y	-1.180	-3.648	-2.958	-2.612	非平稳
X_1	-1.867	-3.648	-2.958	-2.612	非平稳
X_2	-1.761	-3.648	-2.958	-2.612	非平稳
X_3	-1.473	-3.648	-2.958	-2.612	非平稳
X_4	-1.151	-3.648	-2.958	-2.612	非平稳
X_5	-1.655	-3.648	-2.958	-2.612	非平稳
ΔY	-5.558	-3.655	-2.961	-2.613	平稳
ΔX_1	-4.717	-3.655	-2.961	-2.613	平稳
ΔX_2	-4.623	-3.655	-2.961	-2.613	平稳
ΔX_3	-4.593	-3.655	-2.961	-2.613	平稳
ΔX_4	-3.815	-3.655	-2.961	-2.613	平稳
ΔX_5	-4.631	-3.655	-2.961	-2.613	平稳

① 李子奈:《计量经济学》,高等教育出版社,2000年,第269页。

接下来检验被解释变量与解释变量之间是否存在长期稳定的关系,使用 EG 检验法,第一步利用最小二乘法估计回归模型从而得到残差序列;第二步对模型残差进行单位根检验,检验结果显示:由于 $t = -4.691 < -3.627$,表明残差序列在 1% 的显著性水平下拒绝原假设,接受不存在单位根的结论。因此,CCI 与搜索指数之间存在长期稳定的协整关系。因此,本文构建如下模型,见表 3。

在该模型中拟合优度较高,调整 R^2 达到 0.899;然而在进行 Durbin-Watson 检验时发现该模型 DW 值为 1.272,说明模型存在序列相关。为消除该问题,考虑在模型中加入被解释变量 1 阶滞后项。结果表明加入第 1 期的被解释变量后模型变化显著,拟合优度 (Adj R-squared) 达到 0.927, DW 统计量升高至 1.81。因为 DW 检验对存在滞后被解释变量的模型无法检验,考虑使用 Breusch 和 Godfrey 提出的 BG 检验,用于检验高阶序列相关及模型中存在滞后被解释变量的情形。BG 检验原假设是模型不存在序列相关。所以模型以 p 值 0.0001 显著拒绝原假设,即模型存在序列相关。检验结果显示 BG 检验 p 值为 0.6110,说明修改后的模型不存在序列相关的问题。

表 3 回归结果及协整检验

	模型 1	模型 2
截距项	61.822***	44.420***
X_1	0.694*	0.419
X_2	2.566***	1.834***
X_3	1.785***	0.703
X_4	-0.585	-0.320
X_5	-1.886***	-1.234***
Y_{t-1}		0.439***
调整 R^2	0.882	0.913
F 值	54.955	63.726
DW	1.272	1.810
残差平稳性	ADF 值	-4.691
	1% 临界	-3.627
	5% 临界	-2.946
	10% 临界	-2.612
	结论	平稳
结论	协整	协整

(二)模型解释

加入被解释变量 1 期滞后项后的模型回归方程如下,方程表明 CCI 的变化由就业形势,房地产形势,宏观经济变化,家庭经济情况,汽车市场以及消费者对历史经济市场的判断共同决定。

$$Y_t = 44.42 + 0.419X_1 + 1.834X_2 + 0.703X_3 - 0.32X_4 - 1.234X_5 + 0.439Y_{t-1}$$

解释变量 X_1 、 X_2 、 X_3 、 X_4 和 X_5 为搜索变量,反映了消费者对就业市场、房地产市场的关注度,对地区经济景气状况、家庭经济状况的信心,以及对汽车市场的关注情况。

(1)前三个解释变量的系数为正,表明了消费者对这三类搜索指数关注度的提高会引起 CCI 的上升。结合解释变量中包含的搜索关键词进行分析:

在“就业”类别中,保留下来的 6 个关键词对该解释变量贡献大的有“就业服务”“职训局”“劳工局”,市场主体对这类关键词的搜索表现出其对就业市场的看好,即他们愿意主动寻找就业机会,参与职业培训。

在“房地产”类别中,保留下来的 13 个关键词对解释变量贡献大的有“real estate”“奇奇”“房地产”“法拍屋”,其中值得说明的是:“奇奇”这一搜索指数在合成解释变量时的系数为负,这与“奇奇”本身的含义有关,该词代表台湾最大的在地产业类广告平台“奇集集”,该网站集中了丰富的租屋售屋信息,民众的卖屋出租行为可以反映出其并不认为“未来半年是购买耐久性财货的好时机”。

在“宏观经济”类别中,保留下来的 7 个关键词对解释变量贡献大的有“贷款”“银行汇率”“银行利率”,台湾民众对这类关键词的搜索表现出其对经济形势的看好:“贷款”说明资金将导入市场、刺激消费、并协助企业重建资产负债状况。对该关键词的搜索行为表明:企业家将开始接收订单、购买原料、展开生产行为,而对普通消费、投资者来说则表现为其即将开展购买行为。

(2)后两个解释变量的系数为负,表明了消费者对这两类搜索指数关注度的提高会引起 CCI 的下降。在这里进行与上部分相同的分析:在“家庭经济”类别中,保留下来的 19 个关键词对解释变量贡献大的有“台湾人寿”“笔电评比”“产物保险”“富邦产物”,不难发现这些关键词多为保险类资讯信息(富邦产物是台湾第一家民营产物保险公司,为个人和企业提供多种保险类型,如财产险、产险、意外险等),消费者通过购买保险来覆盖未来可能发生的风险,因此消费者对于保险类关键词的搜索行为反应出其对未来经济景气的担忧与信心不足。

在“汽车”类别中,保留下来的 17 个关键词对解释变量贡献大的有“nissan”“honda”“cadillac”“mitsubishi”,关于汽车类关键词搜索量对 CCI 的影响是无法准确定义的,因为人们对这类关键词的搜索既可以表示其对于买车的需求,也可以反映出他们卖车的打算。本文在这部分从负相关的角度来进行解释:由于车贷

已经控制了人们的购车需求,因此该类别关键词的网络搜索量增加反映了大众卖车行为(在网络寻找汽车买家)的上升趋势。

由于时滞期的调整在搜索指数的合成之前,因而无法确定每个解释变量的滞后阶数,但是在所有选取的关键指数中 12 期滞后搜索指数所占比例最大 39 个,11 期滞后的搜索指数 11 个。其它滞后期的搜索指数个数详见表 2。

五个搜索变量的不同系数,表明了五个类别对消费者信心不同程度的影响,这五类指数各自变化 1%,对应消费者信心变化幅度分别为 0.419 个百分点、1.834 个百分点、0.703 个百分点、0.32 个百分点和 1.234 个百分点。这说明消费者对“就业”以及“汽车”的关注程度及占据影响消费者信心指数(CCI)的主要因素。

模型中的为 Y_{t-1} 历史项,第 t 期的 CCI 与第 $t-1$ 期的 CCI 显著正相关,说明了过去一个月的消费者信心对当期消费者看法有着同向程度的影响。

(三)模型预测

为进一步检验网络搜索行为对 CCI 的预测能力,本文将 2010 年 12 月至 2013 年 12 月的数据作为建模数据,对 2014 年 01 月至 2014 年 04 月的 CCI 进行预测,结果显示预测的平均绝对误差为 2.396,均方根误差为 2.749,模型具有良好的预测能力。

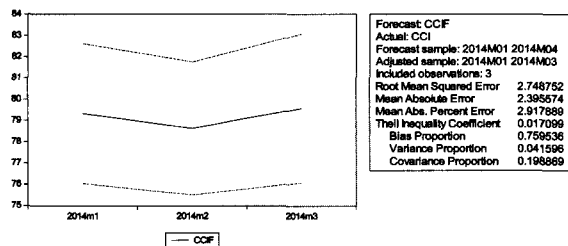


图2 CCI三期预测值

由于模型解释变量中的搜索变量是由 $t-12$ 至 $t-4$ 期搜索指数合成的,历史变量为 $t-1$ 期,因此在获得 $t-1$ 期的所有搜索指数数据时即可预测出第 t 期的 CCI,即能够比中央大学台湾经济发展研究中心公布的数据提前 1 个月左右。基于此,本文认为使用网络搜索合成指标作为反映消费者信心变化的即时指标具有很强的时效性,且为传统的调查方式提供了有效的补充,可以用来帮助对 CCI 进行补充性检测。

在很多经济指标的预测中,相关部门与经济学家关心的问题还包括预测经济指标的结构性变化,即转折点预测能力。传统的预测方法一般都是通过历史数据拟合值来预测未来可能发生的变化,而这样的历史

数据很难判断未来可能发生的不确定突发事件,例如突发的 9·11 恐怖分子袭美事件造成欧美股市重挫,9·16 台湾纳莉台风肆虐使得台北严重淹水引发重大损失,这两事件对台湾经济都造成了重大打击。面对低迷的经济与可能发生的战争多数消费者对未来的景气都不看好。这时基于历史数据的传统预测方法失效,而网络搜索数据可以让我们及时地捕捉到消费者的关注点及兴趣,从而能够实时甚至提前地反映出经济变化的趋势。这为经济指标的拐点预测提供了有效的解决方案。

使用已建立的模型带入搜索数据计算出每期的 CCI 预测值(2010 年 12 月至 2014 年 04 月),图 4 显示 CCI 实际值与通过模型拟合出的预测值的变化曲线,可以看出 CCI 以及搜索数据经历了次较大的三次转折:2011 年年中由骤然上升转为急剧下降的第一次转折,而后在经历了年末的又一次小起伏后在 2012 年末又急剧上升的第二次转折,最后一次转折是在 2013 年初经历的一次波动,而后从年中开始平稳上升。因为事先调整过所有数据的滞后期,所以预测指标与实际指标的曲线应该呈现出近似重合,从图像我们可以看出预测指标的每个转折点都对应了实际指标的每次波动,在第二次波动时甚至达到了完全重合。由此本文认为网络搜索数据对于 CCI 转折点具有一定的预测能力,可以使用关键词的搜索量对 CCI 进行转折点预测。

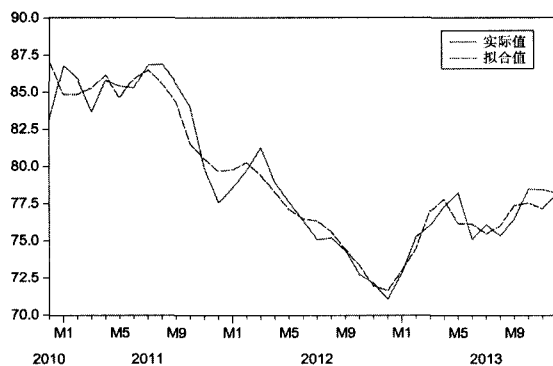


图3 CCI与搜索数据拟合效果图(模型2)

五、结论

实证研究结果表明:基于网络搜索关键词数据建立的 CCI 预测模型其拟合优度达到 0.927,网络搜索关键词数据与 CCI 之间具有长期稳定的协整关系。总体搜索指数的提高会引起 CCI 的上升。模型的预测能力体现在其相对于 CCI 的调查公布有 4 个月的提前期。

台湾辅仁大学对 CCI 的调查以电话访问的方式进

行,共访问台湾地区2400名左右民众。调查时间在每月的19日至23日左右。这样的调查按月进行无疑使工作量变得十分繁重。相比较而言,采用基于网络搜索数据而建立的预测模型来进行CCI调查则只需将模型中所包含的五个类别关键词的Google搜索量进行下载处理,而后带入已建立好的回归方程就可以得出CCI的预测值,这无疑会使调查工作的工作量大大减小。

对比于传统的监测方法,基于网络搜索的CCI模型具有很强的时效性。因为由传统调查方式统计的每月CCI公布时间在每月月末,而通过模型拟合的CCI弥补了传统监测方法一周的信息发布滞后,并且比中央大学台湾经济发展研究中心的数据发布提前1个月左右,由此实现了CCI实时监测功效。同时也使网络搜索数据为传统调查得出的CCI提供有效的补充,加强其准确度。

对比于传统的预测方法,基于网络搜索的CCI模型具有转点预测能力。因为这种预测方式将即时信息带入模型,克服了传统时间序列分析预测、回归分析预测等方法仅依照历史数据来预测未来值的局限性。模型拟合值的转点均与CCI实际值一致,提前期为1个月,故而对CCI即将产生的波动有预测能力。

本文的难点在于关键词的选取、对搜索指标的再分类和对搜索变量的解释。

(1)因为是为人为选取关键词,即使是按照调查CCI时划分的每一类别来搜索词汇,也会存在没有覆盖到的相关类别关键词。这主要是由于关键词的选取存在很大程度上的主观判断,即在Google Trends中输入主观上可能对CCI有强影响力的关键词时,会出现搜索量不足的情况。这说明个人的认知和看法在很多时候并不能代表大众在进行搜索行为时的关注点与兴趣点。因此,在现有条件下只能尽可能多地尝试每一个

类别下可能出现的关键字词,并充分利用Yahoo!奇摩和Google Trends中的相关/热门关键词推荐功能来进行关键词的深度挖掘,最终实现网络式覆盖。该方向在下一步的研究重点就是如何建立一套严谨完善的选词机制,来确保所选关键词的解释能力与词库构成的完整性。

(2)在搜索指标合成方面,因为本文采用的方法是先对所有关键词进行筛选,最后利用主成分分析法对每一类别下所属关键词进行合并。因此如何最优地对筛选后的关键词进行再分类成为了本文的另一难点。如果错误地将本类别的关键词分入他类,就会导致整个类别对CCI解释效果的下降;另外一种错误可能发生于将一类关键词划分为几小类或者将几小类关键词合并为一类,这种行为同样会导致解释不充分的情况出现。

(3)由于形成模型时的各个搜索变量已不再分属于最初查找关键词时的类别设定,而是根据其性质重新划分了类别。因此在解释模型时需要重新理解每个搜索变量对CCI正向或负向影响的背后的真正原因。例如“汽车”类别变量的系数为负,表明了民众对汽车品牌相关关键词搜索量的增大(对汽车关注度的提高)会引起CCI的下降。在一般认知里,民众对汽车品牌关注度的上升可能是由于其购车或换车意图,而购车或换车行为是出于对家庭或个人经济状况的满意判断。按此推论,指标系数应为正,即对汽车关注度的提高会引起CCI的上升。这里便出现了实际模型与解释相悖的情况。因此,为充分解释模型,还需要再继续深入理解经济现象背后的运作机理。

责任编辑:毛伟