·电子商务·

基于互联网搜索信息的预测模型研究
——以余额宝产品需求为例

张爱华，韩怡嘉
（北京邮电大学 经济管理学院，北京 100876）

摘 要：基于百度指数，以余额宝资产规模为例，运用 ARIMA 模型构建引入互联网搜索量的市场需求预测模型，同时建立不含互联网搜索量的预测模型，并将二者进行对比研究，发现包含互联网搜索量数据的预测模型的预测准确度更高，说明互联网搜索量数据包含一些在传统模型中未能考虑到的因素，该变量的引入有利于提高模型的预测准确度，为产品或服务的市场决策提供更加精确的预测结果。

关键词：互联网搜索量；百度指数；余额宝资产规模；预测模型
中图分类号：F713.36；G203；G303 文献标识码：A 文章编号：1008-7729(2015)03-0036-06

一、引 言

搜索引擎作为网民经常使用的工具记录了数以亿计的网络搜索数据。一些国内外最新的研究进展表明，互联网搜索数据与诸多社会、经济行为存在很高的相关关系。基于网络搜索数据的股票预测、房屋销售、失业率、旅游区游客量等的预测都具有较高的准确度，网络数据的及时性能够很好地弥补传统预测方法的滞后性[1-6]。因此可以利用网络搜索数据对社会经济活动进行监控及预测，它能从大量搜索数据中获取新的关联信息，能从复杂的数据中通过数据分析创建合成指数，找出解决问题的途径。

学者们做出很多尝试来分析网民的实际搜索行为所产生的丰富的信息。这些搜索流量信息不仅可以用来预测需求，同时也能运用到其他领域预测未来趋势。Ginsberg et al 曾提出一个模型来评估现有流感的水平，模型分析了谷歌搜索引擎提供的查询信息数据，这个预测结果通常比疾控中心发布的报告早一到两周[7]。而在产品的需求预测领域，Cho et al 发现谷歌趋势能够增强对当前经济活动的预测，例如，在某段时间“房差代理”这个词搜索量的增加，将预示着不久后房屋销售量会增加。所以他们提出将谷歌指数运用到新产品预测的模型中，在自回归方法的基础上加上谷歌指数来预测产品未来的需求，对产品未来需求的预测更加严谨[8]。

二、引入互联网搜索量变量的预测模型建立

1. 不含互联网搜索量的预测模型建立

选取的预测数据是以月为跨度的余额宝资产规模数据，数据的跨度较小，波动较大，不存在明显的趋势。因此模型的建立以依靠自身历史数据预测未来的时间序列预测方法为基础。在时间序列的预测中，对于存在波动的时间序列的预测方法有自回归模型、移动平均模型和自回归移动平均模型等。以自回归模型为例，构建的基本模型如下：

$$y_t = \alpha_{1a} p_{1a} y_{t-1} + p_{2a} y_{t-2} + \varepsilon_t \quad (1)$$

在构建时间序列模型过程中，自回归变量的阶数一般都在 2 阶以内，对于较高阶数的研究，除了

收稿日期：2014-10-23
作者简介：张爱华(1964—)，女，安徽六安人，北京邮电大学经济管理学院教授，硕士生导师，主要研究方向为社交网络、统计、预测和决策支撑。

一些特殊的时间序列模型外，一般在构建模型中应用较少。

在时间序列数据的模型构建时，若要使用 ARIMA 模型，则必须是平稳的时间序列数据，对于一些不平稳的时间序列数据，将其基本模型构建为一个关于时间的回归模型，该回归模型如下:

$$y = \alpha + \beta t + \varepsilon_t \tag{2}$$

其中: y 为产品或服务的市场需求量; t 为时间; ε_t 表示随机变量。

2. 引入互联网搜索量的预测模型建立

在市场需求的预测中，除了在多元回归模型中引入了多个因素外，在时间序列模型中，许多学者还引入了外生变量来提升模型的预测准确度。在早期的时间序列需求预测建模中，许多学者引入了国内生产总值（GDP）、产品价格[8-10]等外生变量，研究发现变量的引入有利于提高模型的预测准确度。随着互联网技术的发展，学者们开始研究互联网中的信息对现实生活的预测作用。在 2006 年，谷歌公司（Google）推出谷歌趋势（Google trends）[11]后，一些学者的注意力开始转向了互联网搜索量，分析互联网中用户在搜索引擎上关于某个关键词的搜索量与现实生活中事物的变化之间的关系。研究发现互联网中用户的搜索量数据与现实生活中物质的需求、自杀率的变化等存在显著的相关关系[12]。

基于以上研究成果，将继续探索互联网搜索量与现实生活中产品的市场需求之间的关系，并在预测模型（1）中引入互联网搜索量，比较其与基本模型的预测效果，建立对比模型如下:

$$y_t = \alpha_{1b} + p_{2b} y_{t-1} + p_{2b} y_{t-2} + \lambda s_t + \varepsilon_t \tag{3}$$

对于某些不平稳的时间序列数据，以模型（2）为基本模型构建对比模型如下:

$$y = \alpha + \beta t + \lambda s + \varepsilon_t \tag{4}$$

其中: y 为产品或服务的市场需求量; t 为时间; s 为该产品或服务的搜索量数据; ε_t 表示随机变量。

三、引入互联网搜索量的余额宝资产规模预测研究

1. 余额宝资产规模数据

余额宝是由第三方支付平台支付宝为个人用户打造的一项余额增值服务。由于具有广大的支付宝用户数和高于银行同期利率近 14 倍的收益作为基础，自 2013 年 6 月推出以来，用户数和资产规模上升迅猛，并且不断深入人们的生活。余额宝资产规模数据如表 1 所示。

表 1　余额宝资产规模数据

年份	2013 年							2014 年		
月份	6 月	7 月	8 月	9 月	10 月	11 月	12 月	1 月	2 月	3 月
资产规模/万元	57	229	402	558	817	1 264	1 853	2 500	4 980	5 413

数据来源: 速途研究院网站 http: //www. sootoo. com/content/483038. shtml，天弘基金微博平台。

2. 余额宝资产规模数据分析

在产品刚开始推广的时期，虽然针对产品做了大量宣传但用户对产品的性能、质量都不够了解。所以大家会选择通过互联网搜索产品信息，为自己的决策提供依据。

从图 1 中可以看出，余额宝资产规模随着百度月关注度的递增呈现上升趋势，二者之间存在明显的相关关系。下面将主要研究基于百度月关注度的余额宝资产规模的预测。

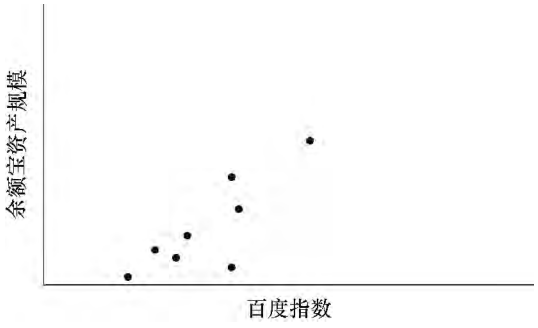

图 1　余额宝资产规模数与百度月关注度散点图

（1）数据平稳性分析

根据图 1 中的散点图可以看出，余额宝整个发展阶段资产规模数据都是呈线性增长的，没有稳定的均值和方差。因此，该时间序列数据不具有平稳性，需要对其进行变换，使之具有平稳性。

（2）自相关与偏自相关分析

表 2 和表 3 的研究结果表明，余额宝资产规模一阶差分后的自相关系数和偏自相关系数每个时期的数值都比较小，都在 0 左右变动。从这些数据中可以进一步看出该时间序列数据具有平稳性，并且在进行时间序列建模时以一期自回归为基础进行时间序列模型的建立。

表 2 余额宝资产规模一阶差分自相关表

滞后	自相关	标准误差	Box-Ljung 统计量		
			值	df	Sig
1	0.083	0.284	0.085	1	0.770
2	0.086	0.266	0.190	2	0.909
3	-0.021	0.246	0.197	3	0.978
4	-0.133	0.225	0.547	4	0.969
5	-0.186	0.201	1.401	5	0.924
6	-0.175	0.174	2.410	6	0.878

表 3 余额宝资产规模一阶差分偏自相关表

滞后	1	2	3	4	5	6
偏自相关	0.083	0.080	-0.034	-0.137	-0.166	-0.137
标准误差	0.333	0.333	0.333	0.333	0.333	0.333

3. 时间序列预测模型分析

（1）不含互联网搜索量的预测模型拟合

从表 4 中可以看出，R^2 的值为 0.861，说明从拟合优度来看，预测模型的拟合结果较优。预测模型的平均 MAPE 为 49.067，表明该模型的预测值和实际值相差较大，模型的拟合准确度有待提升。

表 4 不含互联网搜索量的余额宝资产规模增长模型拟合结果表

拟合统计量	R^2	RMSE	MAPE	MaxAPE	MAE	MaxAE	正态化的 BIC
均值	0.861	778.001	49.067	182.402	430.54	1 885.784	13.802

从表 5 中可以看出，预测模型的系数为 0.079，t 检验值为 0.209，显著性为 0.841，说明该系数的统计检验不是很显著。原因是 2014 年 3 月国家最新颁布的关于限制余额宝单笔支付限额为 1 000 元的政策，影响了余额宝用户对于余额宝的认购力度，直接致使 4 月份余额宝的资产规模增长率下降，增速变缓，影响了预测模型的参数值，导致结果不是很显著。因此，可以得出不含互联网搜索量的预测模型如下:

$$y_t = 0.079 y_{t-1} + v_t \tag{5}$$

其中: y_t 为一阶差分后的余额宝资产规模; v_t 为随机变量。

表 5 不含互联网搜索量的余额宝资产规模增长模型参数估计表

				估计	SE	t 检验值	Sig
余额宝资产规模	无转换	常数		589.700	278.952	2.114	0.072
		AR	滞后 1	0.079	0.378	0.209	0.841
		差分		1			

以此预测模型得出的拟合曲线如图 2 所示。

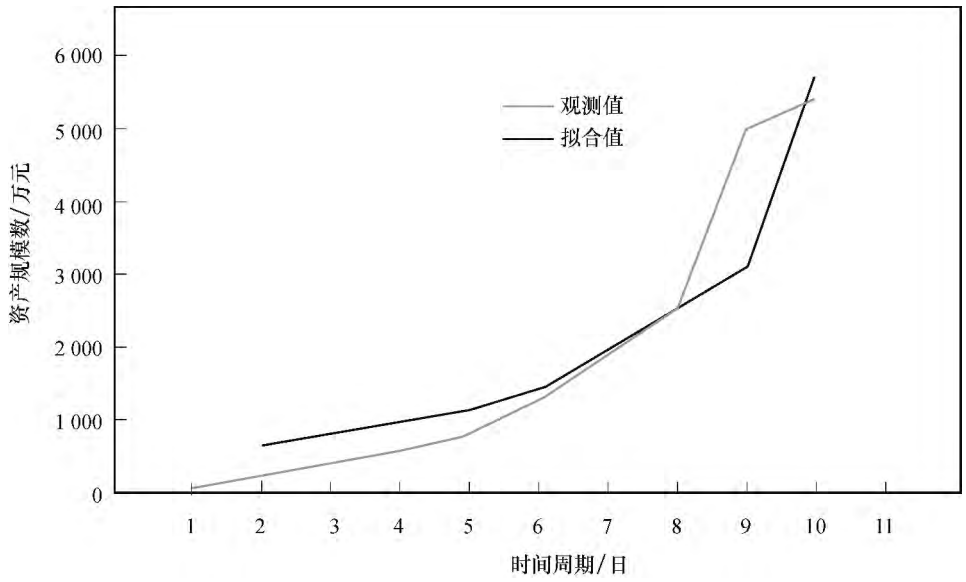

图 2　不含互联网搜索量的预测模型拟合结果图

(2) 引入互联网搜索量的预测模型拟合

表 6 的拟合数据显示，包含互联网搜索量的预测模型的 R^2 为 0.974，说明预测模型的拟合结果和实际值之间的拟合度很高。其中平均误差百分比（MAPE）为 24.954，这也从误差占比角度说明了误差在原始数据中所占的比重比较低。

表 6　引入互联网搜索量的预测模型拟合结果表

拟合统计量	R^2	RMSE	MAPE	MaxAPE	MAE	MaxAE	正态化的 BIC
均值	0.974	365.013	24.954	117.151	208.334	634.097	12.532

从表 7 可以看出包含互联网搜索量的预测模型的参数估计情况，其中自回归系数的检验显著性为 0.097、余额宝月关注度的检验显著性为 0.001。说明模型的系数具有显著性，因此得出包含互联网搜索量的预测模型如下:

$$y_t = -0.907y_{t-1} + 0.001s_t + v_t \tag{6}$$

其中: y_t 为余额宝资产规模一阶差分后的数据; s_t 为当月百度用户关注度; v_t 为随机变量。

表 7　引入互联网搜索量的预测模型参数估计表

				估计	SE	t 检验值	Sig
余额宝资产规模	无转换	常数		−304.253	155.006	−1.963	0.097
		AR	滞后 1	−0.907	0.525	−1.727	0.135
		差分		1			
		分子	滞后 0	0.001	9.035E−0.05	6.148	0.001

同时，该预测模型的拟合结果如图 3 所示。

(3) 模型拟合效果分析

从表 8 中可以发现: 在模型拟合优度（R^2）、平均绝对误差百分比（MAPE）、平均绝对误差

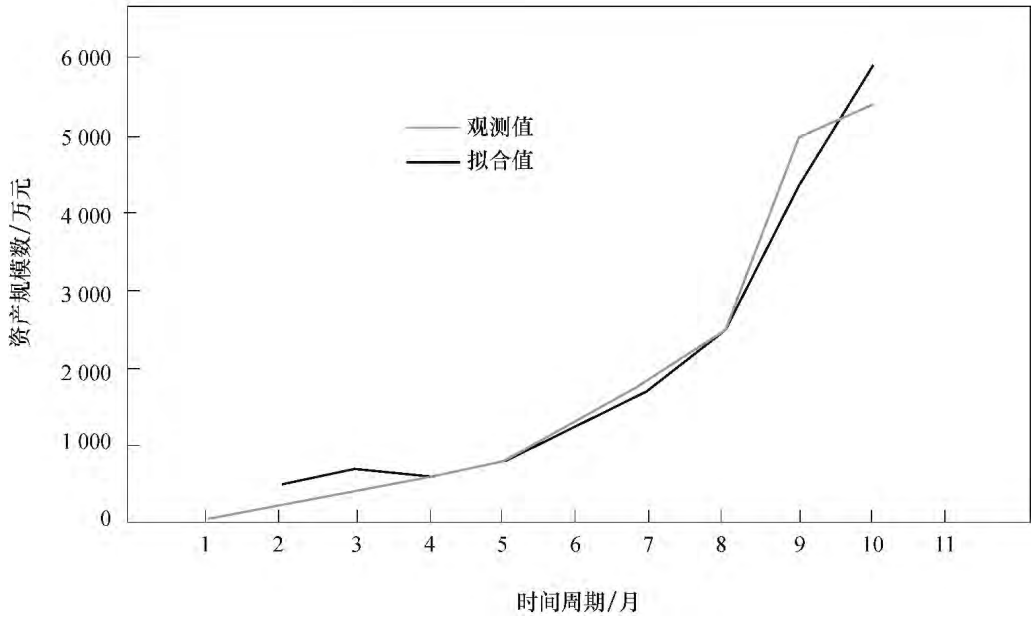

图 3　引入互联网搜索量的预测模型拟合结果图

（MAE）的对比中，包含互联网搜索量的预测模型的统计量都比不包含的预测模型的统计量小，因此包含互联网搜索量的预测模型拟合效果更优。在贝叶斯信息准则（BIC）检验中，包含互联网搜索量的预测模型比不包含的预测模型小 1. 27，这说明变量的引入并不影响模型的有效性。

表 8　包含互联网搜索量的预测模型与不包含互联网搜索量的预测模型对比表

统计量	R^2	MAPE	MAE	正态化 BIC
不含互联网搜索量的预测模型	0. 861	49. 067	430. 540	13. 802
包含互联网搜索量的预测模型	0. 974	24. 954	208. 334	12. 532

综合以上分析，可以得出包含互联网搜索量的余额宝资产规模预测模型比不包含预测模型预测效果更优，预测精度更高。

四、结　论

本研究以国内搜索引擎百度搜索提供的百度指数为基础，以余额宝资产规模数为研究对象，探索互联网搜索量与市场需求之间的关系。研究发现互联网搜索量与产品或服务的市场需求之间存在显著的相关关系。人们可能在购买产品或服务前借助互联网了解该产品或服务的信息，形成了一种潜在的需求。或者在互联网上直接购买该产品，形成该产品的实际销售。

同时通过研究发现，包含互联网搜索量数据的预测模型比不包含的预测模型的预测准确度更高。说明互联网搜索量的数据包含着一些在传统模型中未能考虑到的因素，该变量的引入有利于提高模型的预测准确度，为产品或服务的市场决策提供更加精确的预测。

在研究过程中，涉及的互联网搜索量数据来源于百度指数中的用户关注度。该数据并不是原始的互联网搜索量数据，而是经过加工后的数据，这种数据加工可能会损失原始数据的一些信息，不能完全反映用户在百度搜索上的搜索量，从而对研究结果造成一定的影响。在今后的研究中，可以探寻运用原始的搜索量数据开展研究，避免因为数据加工产生的信息丢失。

参考文献:

[1] 刘颖，吕本富，彭赓．网络搜索对股票市场的预测能力：理论分析与实证检验[J]．经济管理，2011，33(1)：172-180.

[2] Yan Carrière-Swallow，Felipe Labbé. Nowcasting with google trends in a emerging market[R]. Santiago: Central Bank of Chile，2010: 1-19.

[3] Choi Hyunyong，Varian H. predicting initial claims for unemployment benefits[R]. MountainView: Technical Report，Google INC，2009.

[4] Askitas Nikolaos，Zimmermann Klaus F. Google econometrics and unemployment forecast[J]. Applied economics Quarterly，2009，55(2)，: 107-120.

[5] 李山，邱荣旭，陈玲．基于百度指数的旅游景区网络空间关注度：时间分布及其前兆效应[J]．地理与地理信息科学，2008，24(6)：102-107.

[6] 国敏．基于网络搜索技术的游客量预测方法研究——以故宫和泰山景区为例[D]．北京：首都师范大学，2012：1-41

[7] Ginsberg J，Mohebbi M H，Patel R S，et al. Detecting influenza epidemics using search engine querydata[J]. Nature，2009(457)：1012-1014.

[8] 马建忠，邵则夏．核桃、板栗市场需求预测模型的研究[J]．云南林业科技，2000，29(2)：59-62.

[9] 李隽波，孙丽娜．基于多元线性分析的冷链物流需求预测[J]．安徽农业科学，2011，29(11)：6519-6523.

[10] 方庚明．基于多元线性回归的公路客运量发展预测模型[J]．工程与建设，2011，25(2)：164-166.

[11] 程东箭．Google trends 谷歌趋势热力上线[J]．互联网天地，2006(9)：38.

[12] Yang Albert C，Tsai Shi-jen，Huang Norden E，et al. Association of Internet search trends with suicide death in Taipei City，Taiwan[J]. Journal of Affective Disorders，2011(132)：179-184.

Prediction Model Based on Internet Search Information

——Taking Yu'e Bao Product Demand as an Example

ZHANG Ai-hua，HAN Yi-jia

(School of Economics and Management，Beijing University of Posts and Telecommunications，Beijing 100876，China)

Abstract: Based on Baidu index and taking Yu'E Bao asset volume as an example，market demand prediction models containing Internet search volume and excluding Internet search volume are constructed and compared by using ARIMA model. Comparative study shows that prediction model containing search volume has high precision，which means the data contains some factors not considered in traditional model. The introduction of this variable is beneficial to improving the prediction precision of the model，providing more accurate results for the product or service market decision-making.

Key words: Internet search volume；Baidu index；asset size of Yu'e Bao；prediction model