房地产开源舆情指数构建与政策影响研究*

霍琳,尚维,徐山鹰
(中国科学院数学与系统科学研究院,北京 100190)

摘 要 本文选取互联网上大量的新闻数据作为分析对象,根据新闻数量判断公共媒体对房地产市场的关注程度和情感倾向,利用正类、负类新闻数量合成一个反映公共媒体对房地产市场认知的舆情指数。利用2009年11月到2013年3月的数据进行实证检验,证明所构建的舆情指数对商品房销售面积有较强的解释作用,从而能够反映房地产市场的发展变化情况。其有效性来源是房地产政策对市场的影响,事件分析表明房地产政策对房地产市场有较大冲击,冲击从事件日开始逐渐变大,直到第四周后逐渐减弱。本文集成了文本挖掘、计量经济模型和事件分析法,实现了从构建舆情指数到评价其有效性的完整过程。

关键词 社会舆论,情感分析,网络挖掘,房地产市场,事件分析法

中图分类号 C912.63

房地产是中国经济增长的重要引擎,房地产市场的发展情况与个人、企业、政府的信心密切相关。1998年国务院发布《关于进一步深化城镇住房制度改革加快住房建设的通知》,自此中国房地产的市场化运作正式启动。为保持房地产市场持续健康发展,国务院坚定不移地加强房地产市场调控。1998年以来,房地产政策历经多次调整,从稳健到宽松再到紧缩。自2010年起,为了控制房地产市场过热,减少房地产泡沫的风险,国务院开始实行紧缩性房地产政策,限购、限贷政策连续出台,保障房建设投入大量资金。公众急于掌握房地产市场的最新动态,然而传统统计指标无论从可获得性还是时效性上都不能满足公众的需求。

互联网数据,因其开放性和易获得性又称开源数据,指一切可以通过互联网获取的信息,来源包括博客、微博、论坛、电子邮件、新闻报道等。信息形式也是各式各样,有数字的、文本的、图片的、语音的等。海量的开源数据提供了丰富的信息,近年来为越来越多的学者所利用[1,2]。新闻报道作为开源数据的一个主要类别,一方面向公众传达了对真实状况的报道,包括社会、经济、政治等各方面,具有一定的客观性;另一方面新闻报道本身具有一定的感情色彩,如某条关于房地产的新闻,可能表达了对房价进一步上涨的担忧,也可能反映了未来房价有下降的可能。与挖掘网络用户发言中的个人感情色彩一样,挖掘新闻报道背后的态度,也被称为情感分析[3]。

本文通过开源新闻数据分别构建了全国和一线城市的房地产舆情指数,通过实证分析证实舆情指数对全国和一线商品房销售面积有解释作用。该指数能够明显体现出房地产市场的政策,而政策变化对房地产市场销售面积影响也十分显著,因此两者通过政策变化紧密联系在一起。本文以2009年年末到2013年年初的房地产数据为例,检验舆情指数的有效性。

* 基金项目:自然科学基金(71171186,91224006)。
通信作者:尚维,中国科学院数学与系统科学研究院,副研究员。E-mail:shangwei@amss.ac.cn。

1 文献综述

随着房地产经济在我国总体经济体中的重要性日益凸显,人们越来越关注房地产市场的动向。传统的房地产研究中,销售面积和销售价格是学者们长期关注的指标。有的学者考虑到收入因素,将房价收入比作为研究对象[4]。宏观经济指标如国内生产总值、居民消费价格指数等也经常纳入房地产市场分析的指标体系[5,6]。傅劲锋[7]探讨了宏观经济、税收政策、银行抵押贷款政策与房地产的关系。

互联网发展至今,存在大量的零散数据,搜索引擎技术和爬虫技术[8]又为我们提供了从中获取有用信息的方法。Ginsberg、Mohebbi、Patel、Brammer、Smolinski 与 Brilliant[9]利用 Google 搜索趋势成功预测了美国各州的周度流感水平,其预测只有一天延迟,而美国疾病控制和预防中心(US Centers for Disease Control and Prevention,CDC)基于病毒学和门诊数据作同样的预测有一到两周的延迟。Qu、Shang 与 Wang[10]利用新闻搜索结果构建了通胀预警舆情指数,提前两到三个月对 CPI 拐点做出了预警。一些学者从互联网上抓取北京市房屋网上签约套数等数据进行研究,弥补了传统统计指标的不足[11]。

利用传统的统计指标,不仅统计精确性无法确保,且存在诸多缺陷:指标数据不易获得,数据的统计存在时延性,以及统计数据范围不稳定。就房地产市场来说,我国目前的统计指标很不完善,统计粒度较大,不存在一个为大多数学者所承认的景气指数。而国内现有的房地产研究只是从互联网上抓取零散的指标数据,与传统指标相结合。本文利用互联网数据,特别地,利用互联网新闻构建一个反映房地产市场综合情况的指数。

本文整体框架安排如下:第 2 部分介绍利用房地产开源新闻构建舆情指数的方法及具体步骤;第 3 部分介绍检验有效性的方法以及政策影响分析的重要方法;第 4 部分以 2009 年 11 月到 2013 年 3 月的开源新闻为例,利用第 2 部分介绍的方法构建周度房地产开源舆情指数,并利用第 3 部分介绍的方法检验其有效性。实证研究证明,本文构建的指数能够反映房地产市场的景气情况。

2 研究框架及方法

为构建房地产市场舆情指数,研究分为以下几步:①新闻选取。房地产市场是一个周期性市场,其景气周期分为衰退、萧条、复苏、繁荣四个阶段[12],反映房地产市场的景气情况。②确定新闻的情感倾向,是谈论市场上升还是下降。网络新闻的情感倾向与实际市场的景气情况有关。如果房地产市场繁荣,网络新闻中更多地出现与市场"上升"相关的内容;如果情况相反,则网络新闻中更多地出现与市场"下降"相关的内容。③在上述新闻情感倾向基础上,构建反映房地产政策变化的舆情指数。本节详细介绍这三个步骤。

2.1 房地产新闻抓取

在新闻的选取上,本文对比了百度与 Google 新闻搜索,并基于以下原因选择了百度新闻作为新闻来源:一是百度搜索引擎较稳定,在不同时间搜索的结果基本一致,而国内对 Google 的访问时而中断,导致搜索结果中很多为空。二是百度搜索对中文处理有较大优越性,能够更加准确地识别内容。

首先开发网络爬取工具,爬取与主题相关的新闻,并判断新闻的情感倾向。本文爬取百度新闻中包含"房地产"一词的所有新闻。搜索引擎会对关键词进行自动分词,因此在不同文本位置分别包含

“房”和“地产”的新闻也包含在内。按照周度抓取从2009年11月2日到2013年5月12日的新闻，共3亿多条。

2.2 房地产新闻分类

抓取新闻后，通过新闻内容中的情感词将新闻分为正类或负类。如果新闻中包含正的情感词，则认为该新闻的情感倾向为正；如果包含负的情感词，则认为其情感倾向为负。情感词的识别方法如下：以正类情感词为例，人工识别出三个房地产市场上升的阶段，对每个阶段，抓取该阶段的房地产新闻并进行分词，得到分词结果后，去掉对语义分析无作用的停止词(如：的、哎哟、换言之)后，按照词频排序。在排序结果中找出词频高并能够表达“上升”类意思的词作为正类情感词。本文利用刘群、张华平与俞鸿魁[13]提出的基于层叠隐马模型的汉语词法分析系统ICTCLAS，这是目前应用最广的汉语分析系统。

经选取，正类情感词包括：房地产上升，房地产回升，房地产上涨，房地产上行，房地产看好。负类情感词包括：房地产下滑，房地产下跌，房地产回落，房地产下行，房地产看淡，房地产下降。

经过以上分类，得到了每一个时间段的正类房地产新闻数和负类房地产新闻数。部分结果如表1所示。

表1 房地产政策列表

百度-新闻内容搜索结果												
开始时间	结束时间	房地产正类						房地产负类				
		上涨	增加	增长	回升	上升	正类总数	回落	下降	下滑	降价	负类总数
2009-11-2	2009-11-8	335	129	481	46	89	1 080	11	193	13	16	233
2009-11-9	2009-11-15	591	151	797	81	191	1 811	88	296	37	6	427
2009-11-16	2009-11-22	347	140	622	54	69	1 232	19	136	7	5	167
2009-11-23	2009-11-29	361	98	486	19	89	1 053	42	100	9	7	158
2009-11-30	2009-12-6	402	107	745	52	130	1 436	14	80	17	5	116
2009-12-7	2009-12-13	456	375	746	15	121	1 713	13	130	3	2	148
2009-12-14	2009-12-20	1 510	663	808	109	144	3 234	42	384	16	7	449
2009-12-21	2009-12-27	1 180	368	691	18	147	2 404	12	183	17	8	220
2009-12-28	2010-1-3	750	230	484	12	117	1 593	16	148	10	8	182

2.3 舆情指数构建

为计算一定时间段内新闻反映出的舆情指数，可采用正负类舆情的比值取对数来代表整体舆情趋势[14,15]，本文采用公式(1)计算第 t 期的舆情指数：

$$P\ln x_t = \ln \frac{1+\sum_{m=1}^{M} N_{mpt}}{1+\sum_{l=1}^{L} N_{lnt}} \tag{1}$$

其中，N_{mpt} 为第 t 期第 m 个正类关键词相关的新闻数，N_{lnt} 为第 t 期第 l 个负类关键词相关的新闻数。

舆情指数为正，代表公共媒体对房地产市场景气变化趋势的预期为上升，其大小代表预期的强烈程度，越大则公共媒体对房地产市场景气程度上升的预期越强。相似地，舆情指数为负，代表公共媒体对房地产市场景气变化趋势的预期为下降，越小则对房地产市场下降的预警越强。

3 舆情指数有效性与政策影响分析

3.1 有效性分析方法

构建舆情指数后,需要通过实际数据证明指数有效,以进一步应用该指数。本文将商品房销售面积作为被解释变量。选择商品房销售面积主要是基于现有研究[16,17],并且这个指标与舆情指数有较高的相关性。舆情指数作为解释变量引入回归方程,利用回归方程的拟合优度来检验舆情指数的有效性[18,19]。首先建立回归方程。在此回归方程的基础上,增加舆情指数这一解释变量,改进方程。改进方程的拟合优度应高于原方程,且舆情指数作为解释变量应较显著。

3.2 有效性的经济解释

作为一个有效的舆情指数,需要找到其有效性来源。在房地产市场的环境下,本文认为,房地产市场受到国务院出台的房地产政策影响比较显著,房地产市场的销售量和销售价格往往由于新的房地产政策出台而出现比较大的波动。

在新的房地产政策出台前,房地产市场上的交易者通过各种渠道,对政策产生集中预期,同时市场预期与政策实际影响相互作用,也进一步促进政策的公布,加强了房地产市场的波动。公众媒体特别是网络媒体,是市场预期的承载媒介和表现窗口。因此,从理论上来看,基于公众媒体消息构建的舆情指数能够反映房地产交易者的市场预期。从数据上来看,舆情指数所反映出的预警领先于房地产市场的实际波动,因此领先于销售面积这一统计指标。

3.3 政策影响分析方法

为验证房地产市场是否受政策影响,本文采用事件分析法。事件分析法是20世纪70年代在金融领域发展起来的一种方法,其目标是衡量在特定的事件下,公司是否有非正常收益[20,21]。正常收益是指事件没有发生前提下的期望收益,非正常收益是指由事件引起的收益。后来也用来分析某事件对于社会经济生活是否确实有冲击。学者将事件分析法用于许多领域,包括评价反垄断法[22]、分析公司合并的影响[23]等。特别地,已有学者用事件分析法分析"京十二条"房地产政策的影响并取得了非常好的效果[1]。

事件分析法在衡量事件冲击时,首先识别如果事件没有发生的市场表现,然后将其与实际情况对比,将市场的变化作为事件冲击的衡量标准。事件分析法用回归模型来预测事件没有发生的市场表现。

对于房地产市场来说,如果将成交面积作为衡量指标,第 t 期的超额成交面积 AR 定义为

$$\mathrm{AR}_t = x_t - E(x_t) \tag{2}$$

其中,x_t为实际成交面积,$E(x_t)$为模型计算的交易面积期望值。

累计超额成交面积 CAR 定义为

$$\mathrm{CAR}(t_0, t_1) = \sum \mathrm{AR}(t = t_0, t_{0+1}, \cdots, t_1) \tag{3}$$

由于已有研究对单独的政策做分析,说服力较小,本文对多个国务院政策同时进行分析,找出其规律。在对多个事件同时进行事件分析时,一般选取相同的事件窗口长度[24]。

4 实证研究

4.1 指数描述

本文分别构建全国和一线城市的房地产舆情指数。基于互联网数据起始时间的限制，本文所构建舆情指数的时间段为 2009 年 11 月到 2013 年 3 月。所构建的全国和一线城市舆情指数分别如图 1 和图 2 所示。其中，一线城市的选取参考 WIND 资讯数据库[25]中对于房地产市场一线城市的定义，包括北京、上海、广州、深圳四个城市。

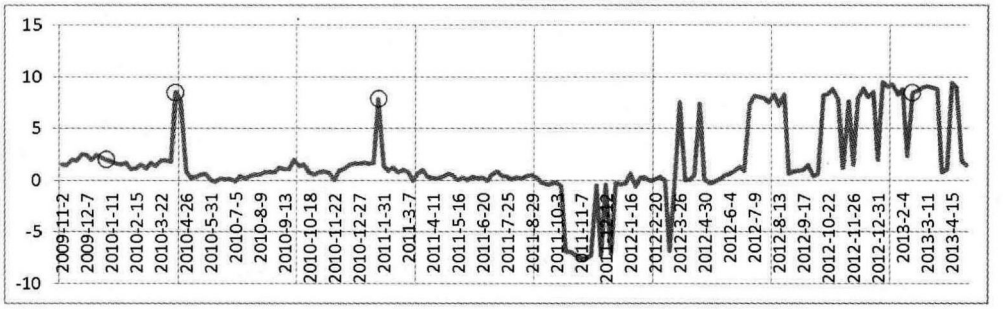

图 1 全国房地产舆情指数

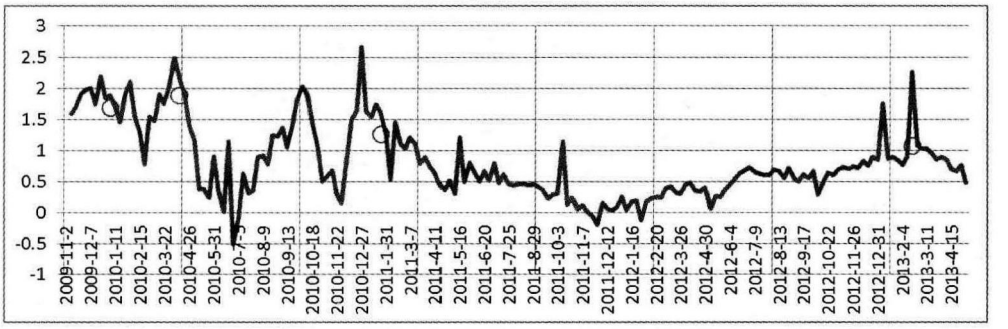

图 2 一线城市房地产舆情指数

全国房地产正、负类新闻的判断如 2.2 节所述。对于一线新闻，除 2.2 节所述规则外，新闻内容中出现“北京”、“上海”、“广州”、“深圳”、“一线”五个词中的任意一个，即认为该新闻谈论的是一线城市房地产状况。

舆情指数表现了公众对于房地产政策变动的关注程度。观察所构建的舆情指数发现，在国务院政策出台时，指数波动较大。图 1 和图 2 中红圈标注的是国务院政策出台的时间点。由图中看出，在政策出台时，全国房地产舆情指数呈现明显波动；一线城市房地产舆情指数波动的出现明显领先于政策出台时间，显示出一线城市的公众媒体对政策更加关注和敏感。表 2 列出了本文所关注的政策及相应时间段的舆情指数。表 2 中，全国舆情指数前后三周是指政策出台当周向前后各推一周的均值，前后十周是指政策出台当周向前推四周，向后推五周的均值。前者舆情指数明显大于后者。公众媒体对一线城市房地产市场更关注，同时政策的实施也从一线城市开始，因此舆情指数开始上升较早。我们用政策出台前后三周一线城市舆情指数的均值与前后十周的均值比较，前者均大于后者。

表2 房地产政策列表

有关政策名称	全国舆情指数		一线城市舆情指数		政策日	事件日
	前后三周	前后十周	前三周	前后十周		
关于促进房地产市场平稳健康发展的通知（“国十一条”）	2.00	1.84	1.98	1.84	2010-1-7	2009-12-27
关于坚决遏制部分城市房价过快上涨的通知（“国十条”）	5.99	1.55	2.22	1.55	2010-4-17	2010-4-18
关于进一步做好房地产市场调控工作有关问题的通知（“国八条”）	3.63	1.46	1.95	1.56	2011-1-26	2011-1-30
关于继续做好房地产市场调控工作的通知（“国五条”）	6.49	1.07	1.46	1.07	2013-2-20	2013-2-3

4.2 评价模型

在房地产舆情指数构造的基础上，进一步，我们希望验证舆情指数对商品房销售面积这一传统统计指标的解释作用。首先引入房地产预测的经典基准模型，然后将舆情指数加入解释变量进行比较。选取2009年11月到2013年3月的月度数据，除舆情指数外的数据都来源于Wind资讯数据库[25]。

基于现有文献研究以及不同变量的代表性、数据的可获得性，本文基于以下原因选取解释变量：广义货币量M2越大，市场上投入流通的货币越多，会在很大程度上提高各种物价，特别是投资品价格。工业增加值一方面可以较好地反映我国国民经济状况，与GDP存在较好的对应关系；另一方面，工业增加值可以比较全面反映工业和建筑业景气状况，水泥、钢材、机械制造等都能够得到较好的反映，这些行业与房地产业的发展密不可分。房地产开发投资完成额反映了房地产自身以及对其相关行业的经济贡献。股票市场交易量代表金融市场的繁荣程度，股票市场交易量越大，金融市场越繁荣，消费者倾向于扩大消费，相应会对房地产市场有拉动作用；相反地，在股票市场不景气时，股票市场交易量较小，市场中的资金流量减少，房地产市场中投资和投机资金也会减少，在一定程度上压低房地产价格和销量[26,27]。

首先，建立如下全国商品房销售面积的回归模型：

$$\mathrm{AREA} = \alpha + \beta_1 \mathrm{D(IND)}(-1) + \beta_2 \mathrm{D(INV)}(-1) + \beta_3 \mathrm{D(M2)}(-2) + \beta_4 \mathrm{LN(ST)}(-1) \quad (4)$$

其中，AREA为商品房销售面积，IND为工业增加值当月同比，INV为房地产开发投资完成额累计同比，M2为货币供应量当月同比，ST为上证和深证所A股成交金额总和。静态回归分析详细结果见表3模型一。该模型修正R方为0.44，模型较显著。

为了将全国房地产舆情指数(SENT)加入模型中，首先将其与商品房销售面积做相关性分析，二者同期相关，相关系数为0.86，二者为强相关关系。协整检验结果表明二者存在长期均衡关系。在满足以上条件的基础上，建立如下回归模型：

$$\begin{aligned}\mathrm{AREA} = {} & \alpha + \beta_1 \mathrm{D(IND)}(-1) + \beta_2 \mathrm{D(INV)}(-1) + \beta_3 \mathrm{D(M2)}(-2) \\ & + \beta_4 \mathrm{D(ST)}(-1) + \beta_5 \mathrm{SENT}(-1)\end{aligned} \quad (5)$$

引入舆情指数后，回归模型的拟合优度有所提升，修正R方提高到0.47，并且舆情指数对商品房销售面积的解释作用较显著。详细结果见表3模型二。

相似地，建立一线城市商品房销售面积的回归模型，模型结构如下：

$$\begin{aligned}\mathrm{AREA\_L1} = {} & \alpha + \beta_1 \mathrm{D(IND\_L1)} + \beta_2 \mathrm{D(INV\_L1)}(-6) \\ & + \beta_3 \mathrm{D(M2)}(-2) + \beta_4 \mathrm{LN(ST)}(-1)\end{aligned} \quad (6)$$

指标名称中 L1 后缀代表一线城市。详细结果见表 3 模型三。该模型修正 R 方为 0.4。

对一线城市房地产舆情指数与商品房销售面积做相关性分析，销售面积相对舆情指数滞后两阶时，相关系数为 0.79，为强相关。协整检验结果表明二者存在长期均衡关系。将一线城市房地产舆情指数引入解释变量后，建立回归模型，结构如下：

$$\begin{aligned} AREA\_L1 = {} & \alpha + \beta_1 D(IND\_L1) + \beta_2 D(INV\_L1)(-6) + \beta_3 D(M2)(-2) \\ & + \beta_4 LN(ST)(-1) + \beta_5 SENT\_L1 \end{aligned} \tag{7}$$

以上模型修正 R 方由 0.4 提高到 0.55，有较大幅度提升，并且一线城市房地产舆情指数(SENT_L1)的解释作用十分显著。回归详细结果参见表 3 模型四。

表 3 商品房销售面积回归结果

全国商品房销售面积回归结果(LN(AREA))				一线城市商品房销售面积回归结果(LN(AREA_L1))			
模型一		模型二		模型三		模型四	
C	4.73(0.69**)	C	4.01(0.77**)	C	3.46(0.64**)	C	3.52(0.56**)
D(IND)(−1)	0.06(0.02*)	D(IND)(−1)	0.05(0.02*)	D(IND_L1)	0.05(0.03*)	D(IND_L1)	0.03(0.03)
D(INV)(−1)	0.01(0.01)	D(INV)(−1)	0.00(0.01)	D(INV_L1)(−6)	−0.02(0.01)	D(INV_L1)(−6)	−0.02(0.01*)
D(M2)(−2)	0.18(0.04**)	D(M2)(−2)	0.14(0.05**)	D(M2)(−2)	0.14(0.04**)	D(M2)(−2)	0.14(0.04)
LN(ST)(−1)	0.15(0.06*)	LN(ST)(−1)	0.20(0.06**)	LN(ST)(−1)	0.20(0.06**)	LN(ST)(−1)	0.18(0.05**)
		SENT(−1)	0.03(0.01*)			**SENT_L1**	**0.27(0.07**)**
R 方	0.50	R 方	0.54	R 方	0.46	R 方	0.60
修正 R 方	0.44	修正 R 方	0.47	修正 R 方	0.4	修正 R 方	0.55
标准误差	0.31	标准误差	0.29	标准误差	0.29	标准误差	0.25
F 统计量	9.01	F 统计量	8.05	F 统计量	7.63	F 统计量	10.60

注：变量后的单元格中，第一个数为回归系数，括号中为标准差，* 表示 p 值小于 0.1，** 表示 p 值小于 0.01。

对比全国和一线城市的房地产舆情指数，后者更显著，对销售面积的解释力度更大，进一步说明公众媒体对一线城市的房地产市场更关注。并且在全国中，舆情指数作为商品房销售面积的解释变量滞后一期，而一线城市舆情指数没有滞后，说明一线城市的商品房销售面积对政策更敏感。

4.3 政策影响分析

为验证房地产市场是否受政策影响，本文选取 2010 年年初至今的四条国务院政策作为分析对象。之所以仅以四条国务院政策作为所有房地产政策的代表，一是因为政策密度太大，作用时间上有重叠，无法单独分析；二是因为其他部委的房地产政策都可以归纳为某一国务院政策出台的细则，国务院政策出台后影响已经显现，相比而言，其他部委的政策细则影响较小。本文研究的政策见表 2。

分别对每个政策进行事件分析，原假设均为：该政策对商品房销售面积没有影响；政策日为政策公布或实施日；事件日由对数据的观察得出，定为数据有明显波动的一周；估计窗口为由事件日向前推 72 周；事件窗口为事件日所在周开始的 8 周。

采用 ARIMA 模型预测事件窗口中交易面积的期望值 $E(x_t)$。为更好地展示房地产政策的影响，定义政策影响为：$\frac{\text{实际销售面积}-\text{预期销售面积}}{\text{预期销售面积}} \times 100\%$。图 3 显示在 8 周事件窗口中，每一周的政策影响。由图可以看出，政策的影响在事件窗口开始时逐渐变大，第 4 周后开始变小，政策影响逐渐变得不再显著。

由 4.1 节对舆情指数的分析看出，政策出台与舆情指数的波动具有紧密联系。舆情指数的波动

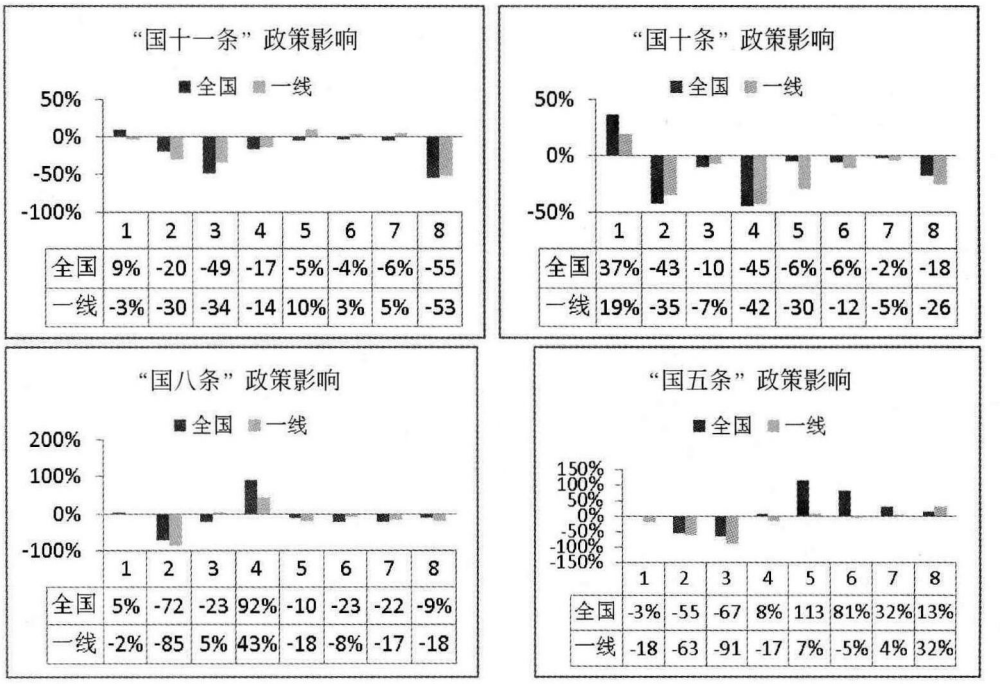

图 3 房地产政策对商品房销售面积影响

受对政策预期的影响，与政策的出台在同一期或略提前。而同样受政策影响，销售面积的波动一般从第二期开始出现。房地产政策这一因素将舆情指数和销售面积联系在一起。

5 结论

本文开发了网页爬取工具，通过百度新闻搜索引擎，爬取一定时间段内、与全国或者一线城市房地产相关的新闻数。通过在新闻内容中识别正负类情感词，判断一条新闻的情感倾向。以正负类新闻数为基础，构建了全国和一线城市房地产舆情指数。与单纯的新闻数相比，舆情指数能够综合反映房地产市场的舆情现状。互联网新闻易获得，并消除了传统指标数据时间上的延迟，以及指标数据只在较大地理范围内聚集的缺点。通过有效性分析验证，在数据量很大时，本文构建的舆情指数对商品房销售面积有解释作用，进而能够反映房地产市场的运行状况。同时，商品房销售面积受房地产政策出台影响，出现较大波动，而舆情指数在对政策的预期作用下较早出现同样的波动，使得舆情指数能够领先于销售面积，反映出房地产市场的波动。

本文提出的方法利用互联网数据构建情感指数，用于描述房地产市场的综合状况。此方法不仅适用于房地产市场，也适用于公众媒体或民众关注的其他行业，如家电市场等。将本文的方法稍加修改即可用于构建其他行业的情感指数，十分有效又简单易行。

本研究中抓取的新闻无法排除对房地产市场过去的评论和未来的预期，也无法排除对国外房地产市场的评价。由于数据可获得性的限制，本文没有分析舆情指数对房地产销售价格的预测作用，以及对二线城市各指标的预测作用。进一步的研究将从两方面展开，一方面是从文本分析的角度，对新闻文本作深入分析处理，从语义角度排除与研究目的关联较少的文本，对文本进行更精确的分类；另一方面是从房地产行业指标的角度，引入其他两个房地产指标——房地产价格和房地产销售量，分析舆情指数对这两个指标的解释作用，以进一步证明其有效性来源，以及与房地产政策的关系。

参考文献

[1] Montes-y-Gomez M, Gelbukh A, Lopez-Lopez A. Mining the news: trends, associations, and deviations[J]. Computacion Sistemas,2010,5(1): 14-24.

[2] Wu L,Brynjolfsson E. The future of prediction: How Google searches foreshadow housing prices and sales[J]. Available at SSRN 2009.

[3] Wiebe J M, Bruce R F, O' Hara T P. Development and use of a gold-standard data set for subjectivity classifications[C]. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics,1999: 246-253.

[4] 薛莉苇,赵晓军,许健.房价收入比影响因素分析[J].浙江社会科学,2010,(3): 17-20.

[5] 沈悦,刘洪玉.房地产价格与宏观经济指标关系的研究[J].价格理论与实践,2002,(8): 20-22.

[6] 闫妍,许伟,部慧,宋洋,张文,袁宏,汪寿阳.基于 TEI@I 方法论的房价预测方法[J].系统工程理论与实践,2007,(7): 1-9.

[7] 傅劲锋.房地产价格理论与实证研究[D].长春: 吉林大学,2007.

[8] Manku,Gurmeet S,Jain A,Das S A. Detecting near-duplicates for web crawling[C]. In Proceedings of the 16th International Conference on World Wide Web. ACM,2007: 141-150.

[9] Ginsberg J,Mohebbi M H,Patel R S,Brammer L,Smolinski M S,Brilliant Larry. Detecting influenza epidemics using search engine query data[J]. Nature,2008,457(7232): 1012-1014.

[10] Qu Y,Shang W,Wang S Y. A Price sentiment index for macroeconomic early warning[C]. The Sixth China Summer Workshop on Information Management (CSWIM 2012),2012: 158-164.

[11] 郭琨,崔啸,王珏,汪寿阳,成思危."京十二条"房地产调控政策的影响——基于 TEI@I 方法论[J].管理科学学报,2012,15(4): 4-11.

[12] 朱骏.基于时间序列与人工神经网络的房地产周期识别[J].清华大学学报(自然科学版),2006,46(6): 781-784.

[13] 刘群,张华平,俞鸿魁.基于层次隐马模型的汉语词法分析[J].计算机研究与发展,2004,41(8): 1421-1429.

[14] Antweiler W,Frank M Z. Is all that talk just noise? The information content of internet stock message boards [J]. Journal of Finance,2004,59(3): 1259-1294.

[15] Qu Y,Shang W,Wang S Y. Webpage Mining for Inflation Emergency Early Warning. In Web-Age Information Management[A]. Springer Berlin Heidelberg,2013: 211-222.

[16] 邱强,万海远.我国房地产业的周期运行特征[J].统计与决策,2007,22: 79-82.

[17] 师应来,王平.房地产预警指标体系及综合预警方法研究[J].统计研究,2011,28(11): 16-21.

[18] 游家兴,吴静.沉默的螺旋: 媒体情绪与资产误定价[J].经济研究,2012(07).

[19] Tetlock P C. Giving content to investor sentiment: The role of media in the stock market[J]. Journal of Finance, 2007,62: 1139-1168.

[20] Fama E F. Efficient capital markets: A review of theory and empirical work[J]. Finance,1970,25(2): 383-417.

[21] Pamela P P. Event studies: A review of issues and methodology [J]. Quarterly Journal of Business and Economics,1989,28(3): 36-66.

[22] Duso T,Gugler K,Yurtoglu B. EU merger remedies: An empirical assessment[A]. Contributions to Economic Analysis. Emerald Group Publishing Limited. 2007,282: 302-348.

[23] Duso T,Gugler K,Yurtoglu B. Is the event study methodology useful for merger analysis? A comparison of stock market and accounting data[J]. International Review of Law and Economics,2010,30(2): 186-192.

[24] Bhattacharya U,Daouk H,Jorgenson B,et al. When an event is not an event: The curious case of an emerging market[J]. Journal of Financial Economics,2000,5: 69-101.

[25] Wind资讯[EB/OL]. http://www.wind.com.cn/,2013.

[26] 王洋.房地产调控的宏观视角[J].上海经济研究,2005,10: 3-10.

[27] 徐国祥,王芳.我国房地产市场与股票市场周期波动的关联性探讨[J].经济管理,2012,34(02): 133-141.

Research on Public Opinion on Real Estate and the Policy Influence

HUO Lin, SHANG Wei, XU Shanying

(Academy of Mathematics and System Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract This research uses the large amount of Internet news as the subject and analyzes how much the public media are focusing on the market and what opinion they have, according to the amount of news. A real estate sentiment index was constructed using the positive and negative amount of news. The empirical study is conducted using data from November 2009 to March 2013, which proves that the sentiment index has strong explanation to sales area of commodity houses, therefore can reflect the state of real estate market. This is because the real estate policies have impact on the market. And the impact is proved by the event analysis. The impact begins from the event day and grows gradually until the fourth week. This paper integrates text mining, econometric models and event analysis techniques, and realized the process from sentiment index construction to its evaluation.

Key words Public opinion, Sentiment analysis, Internet mining, Real estate market, Event study

作者简介

霍琳(1985—),女,中国科学院数学与系统科学研究院,博士研究生。研究方向为决策支持系统、数据挖掘、多主体仿真。E-mail: huolin@amss.ac.cn。

尚维(1978—),女,中国科学院数学与系统科学研究院,副研究员。研究方向包括电子商务、谈判支持系统、宏观经济预警系统、群体决策理论、决策支持系统、信息系统设计与开发、信息系统经济学。E-mail: shangwei@amss.ac.cn。

徐山鹰(1951—),男,中国科学院数学与系统科学研究院主任,高级工程师,博士研究生导师。研究方向包括决策支持系统理论与方法、宏观经济信息系统设计与开发等。E-mail: xsy@iss.ac.cn。