


网络数据在CPI中的应用

中国国家统计局
城市社会经济调查司流通消费价格处

孙易冰

目录

1 研究背景

2 网络爬虫技术及获取的数据特征

3 基于网络爬虫的价格指数编制模型及实证

4 优点和问题

研究背景

必要性

- 近年来，B2C型（www.amazon.com、www.jd.com）和C2C型（www.ebay.com、www.taobao.com）电子商务网站发展迅速，给个人消费行为带来了巨变。
- 网络中本身存在的海量数据也提示官方统计部门将其应用到CPI统计中。

官方统计部门存在的困难

- 官方统计部门（外部数据获取者）和电子商务公司（内部数据拥有者）相比，难以获得真实成交价格，真实销量等数据。
- 难以确定网络销售行为的真实性。
- 难以确定电子商务与传统零售业之间销量的占比。

官方CPI部门仍以现场采价为主

中国CPI现行工作模式

- 采价方式：采价员定期现场采价。
- 权数获取：主要来自13.3万户城乡居民住户收支调查。
- 计算方法：链式拉式公式。
- 质量调整：国际推荐的主要方法。

目录

1 研究背景

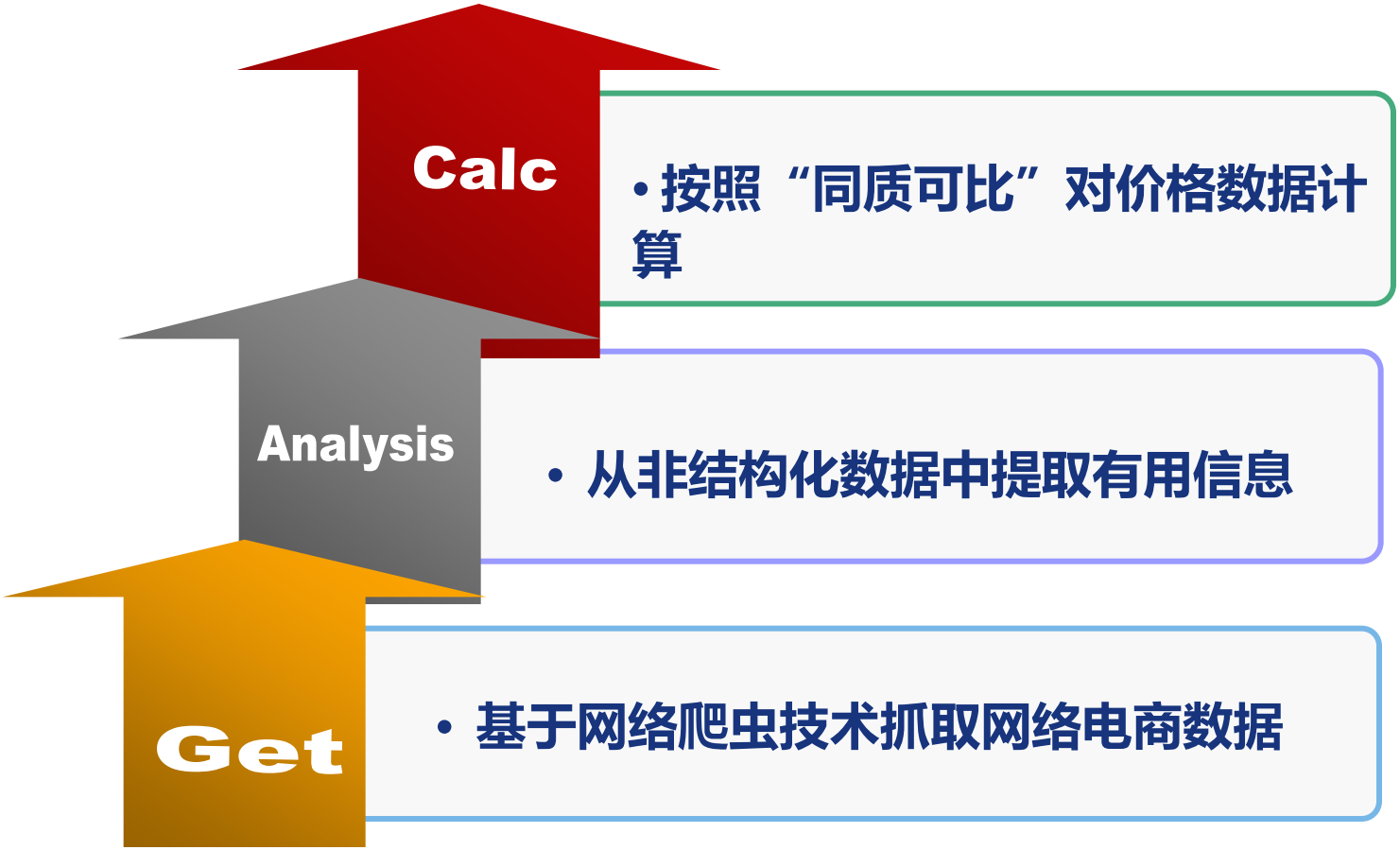
2 网络爬虫及获取的数据特征

3 基于网络爬虫的价格指数编制模型及实证

4 优点和问题

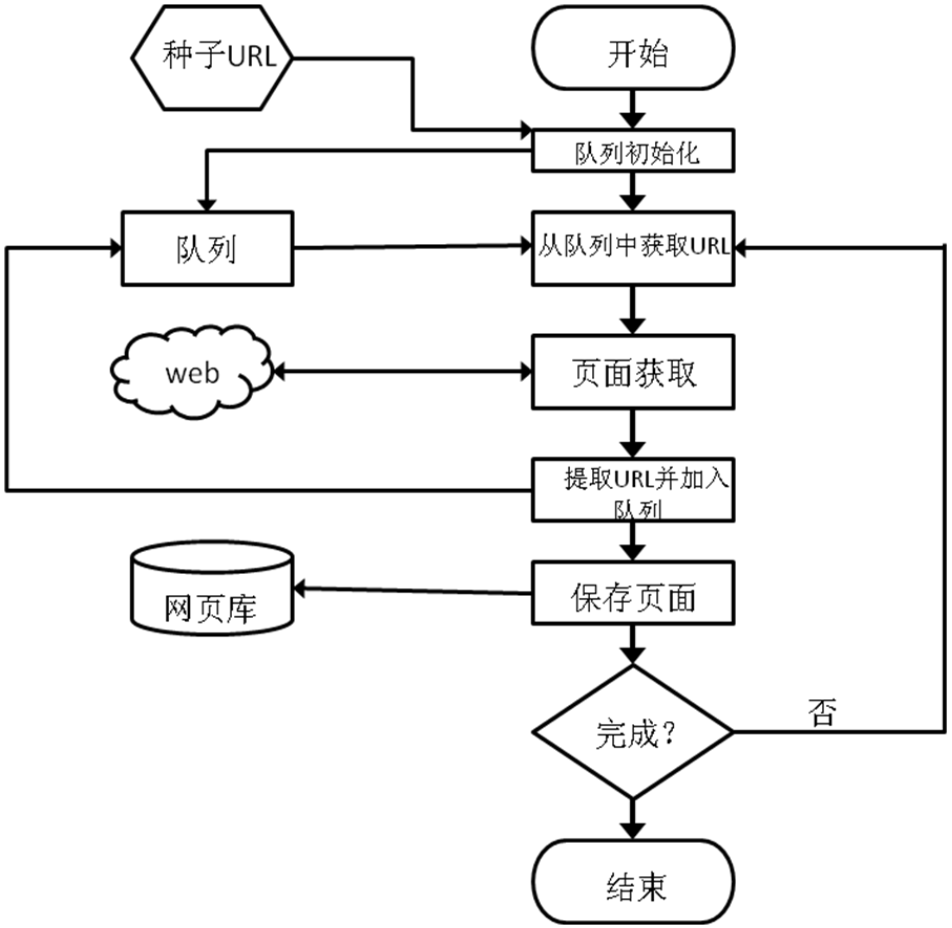
网络爬虫及获取的数据特征

外部数据获取者：基于网络爬虫的工作流程



网络爬虫：按规则对目标网站内容进行遍历，并按照指定格式提取出有用信息。

网络爬虫的基本情况



在Linux环境下采用Perl编写网络爬虫程序。以京东商城（www.jd.com）手机栏目为例，日数据量在200MB左右。

由于电商通常采用HTML类网页格式发布信息，可应用正则表达式等技术对爬虫抓取的页面文件进行解析. 2014年5-9月期间提取出30多万条信息。

案例分析

京东商城手机索引页面

京东商城手机单品页面

提取分析

采集的价格样本

序号	手机型号	价格（元）	有无货	好评	中评	差评
1	三星 B309i	168.00	有	26939	1415	428
2	iPhone4S 8GB版	2448.00	有	54243	2453	1822
3	Galaxy S4 I9300	1899.00	有	39131	1873	982

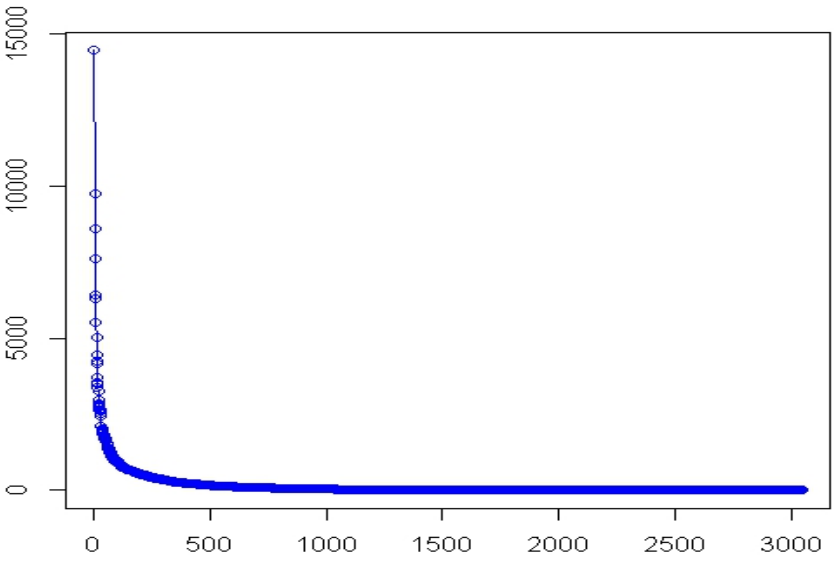
两种方式获得数据的区别

两种方式获得数据的区别

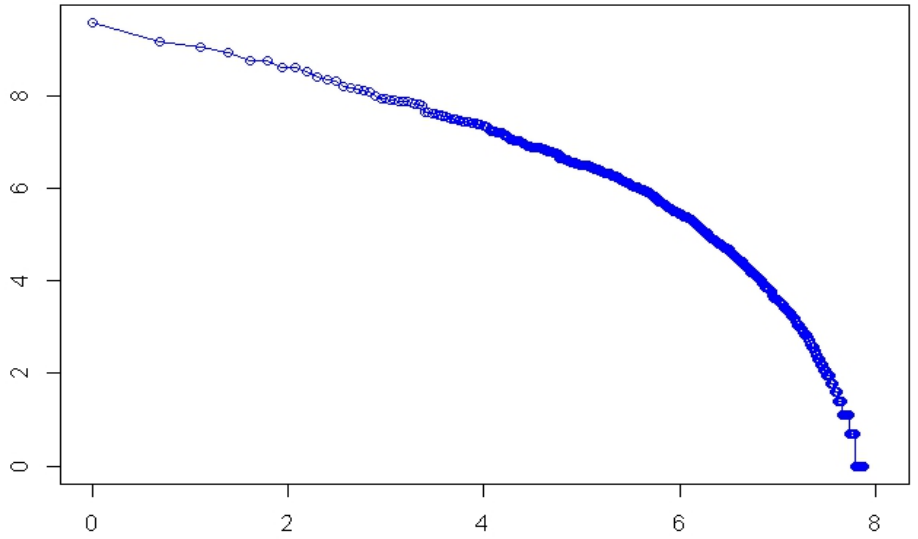
	爬虫采价方式	人工采价方式
采价频率	每天定时采价价格	按照固定时间（3-5天）间隔
采价数量	商品数量丰富。单个电商网站销售手机数量可以达到3000款左右。	按照代表性原则选取固定数量（5-10款左右）的手机参与指数计算。
价格趋势	日度算术平均价格波动剧烈。	价格趋势相对平稳。
促销行为	电商经常采取买赠、秒杀等活动进行促销，难以判别真实价格。	通过现场判断促销行为，识别价格是否为真实成交价格。

通过聚类分析等技术分析数据，电商表现出和传统零售商截然不同的销售策略（往往从库存角度出发，进行一些试探性销售）。需要科学处理秒杀、限时降价、买赠等销售行为。

一段时间内的评论数呈现出明显的长尾分布特征



正常坐标下一段时间内评论数与排序的关系



双对数坐标下一段时间内评论数与排序的关系

通过数据分析，发现电商一段时间内的评论数呈幂律分布（长尾分布）。资料显示，亚马逊公司的商品销量也呈幂律分布。有理由认为电商一段时间内的评论数可近似作为精确销量的判断依据。

网络爬虫技术面临的困难

网络电商设计的页面样式复杂，信息高度冗余，有用信息仅几百个字节。部分网站通过设计价格标签图片等手段防止第三方机构获取数据。

电商经常通过网站改版等技术手段吸引客流。由于爬虫技术高度依赖于文本解析策略，经常需要跟随网站的新版面设计新的解析策略。

由于网站安全策略等原因，获取数据存在困难。

目录

1 研究背景

2 网络爬虫技术及获取的数据特征

3 基于网络爬虫的价格指数编制模型及实证

4 优点和问题

基本分类商品的日度价格指数模型

为与现行方法保持一致，采取国际官方统计部门常用的Jevons公式对基本分类价格指数（低级指数）进行计算，计算相邻两天有价格商品的价格环比波动情况，依此计算月度、年度基本分类商品价格指数，并按照链式拉式公式进行指数汇总。

方案一：将评论数排名前5的手机纳入计算

方案二：将全体手机纳入计算

3种指数计算结果的对比（上个月=100）

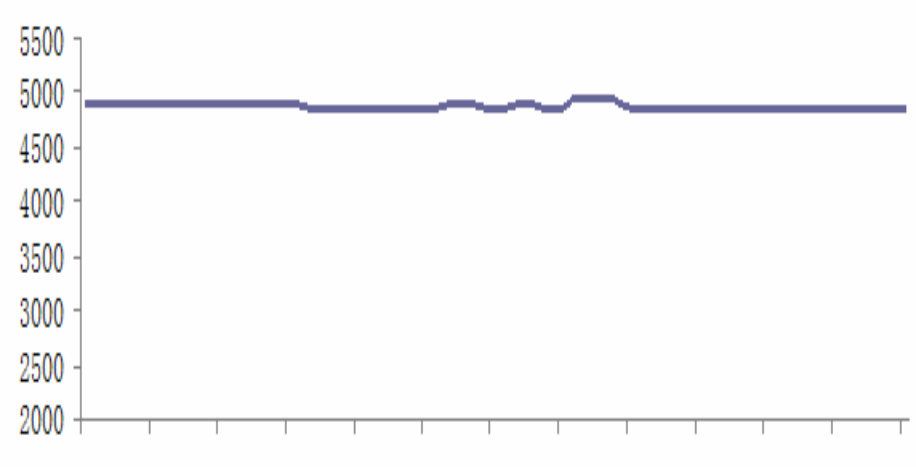
	方案一环比	方案二环比	中国CPI中手机基本分类价格环比指数
2014年6月	99.0	95.9	99.8
2014年7月	99.6	97.7	99.9
2014年8月	99.7	96.0	99.6
2014年9月	99.5	96.5	99.5

基于网络爬虫计算价格指数基本可行

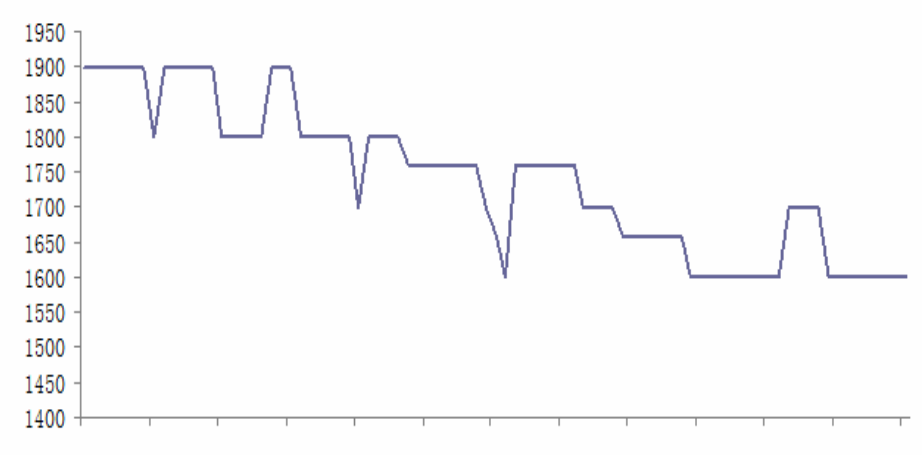
1.基于网络爬虫计算的基本分类价格指数与官方数据基本接近。

2.为什么方案二的数据明显低于方案一和官方数据？

从单品手机价格分析，深受消费者欢迎的手机价格长期坚挺，而其他小品牌手机则通过快速降价意图抢占市场。在不考虑代表性的等权模型下,计算全体手机价格将放大某些产品的降价效应，导致数据失真。



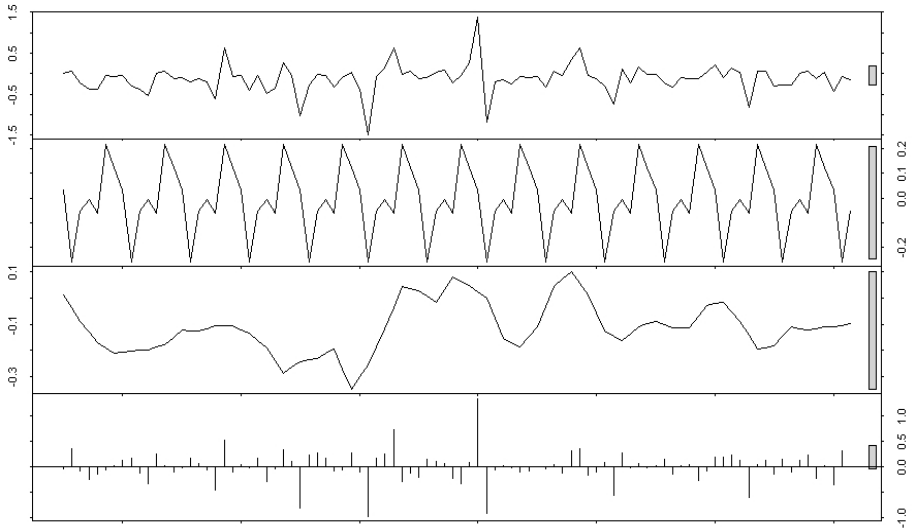
2014年5-9月 iPhone 5S 价格走势



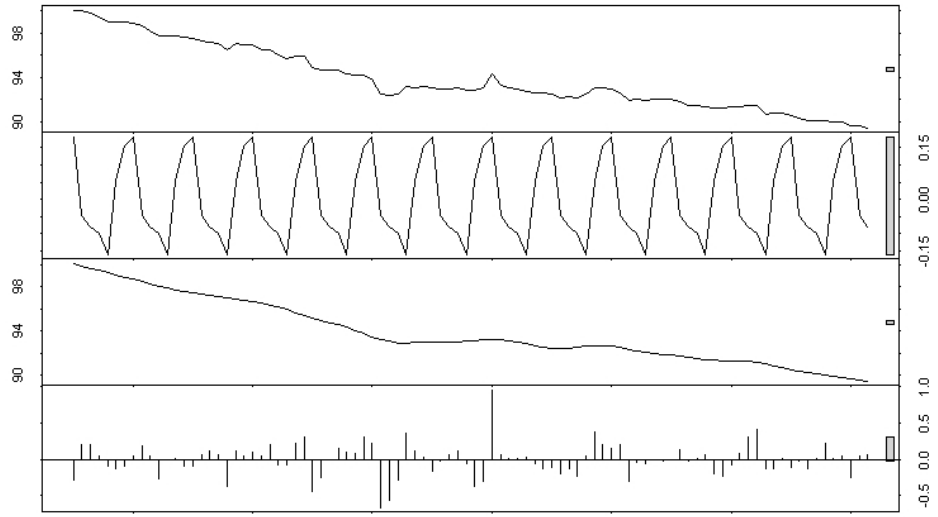
2014年5-9月 HTC D816W 价格走势

从更小的时间尺度观察电商的销售行为及价格波动特征

季节调整（一周）后方案二的数据



方案二的日度价格环比指数STL分解（7天）从上往下分别为：日度环比指数，季节性时间序列、趋势时间序列、剩余项。



方案二的日度价格定基指数STL分解（7天）从上往下分别为：日度定基指数，季节性时间序列、趋势时间序列、剩余项。

1. 从方案二的日度价格环比指数（可以认为是当天电商的促销力度）看，和传统的零售业“周末效应”类似，电商在一周内有明显的周期性降价促销行为（周中降价，周末涨价），这也和从原始数据的分析结果一致。

2. 从方案一和方案二的定基价格指数都可以看出手机价格长期走低的情况。

目录

1 研究背景

2 网络爬虫技术及获取的数据特征

3 基于网络爬虫的价格指数编制模型及实证

4 优点和问题

优点

- 官方统计部门可通过爬虫技术拓展采价方式。
- 和人工现场采价相比较，爬虫技术省时省力，可以实现全面统计。
- 可以从更小的时间尺度观察价格波动。

问题

如何辨别真实成交？

- 互联网产品的销售存在退货现象，极端情况下，部分月份的实际销售额可能为负值，不具备统计学上的代表意义。

如何解决代表性问题？

- 通用的初级指数计算方法都是等权方法。将全部规格品纳入计算必须采用排除选样方法，否则将导致代表性缺失，数据失真。

谢谢！