Application of network data in CPI

YiBing Sun

Department of City Statistics, NBS

Outline

1 Backgound

2 Web crawler and the Data characteristics

3 The price index model based on the web crawler and the empirical

4 Advantages and disadvantages

Background

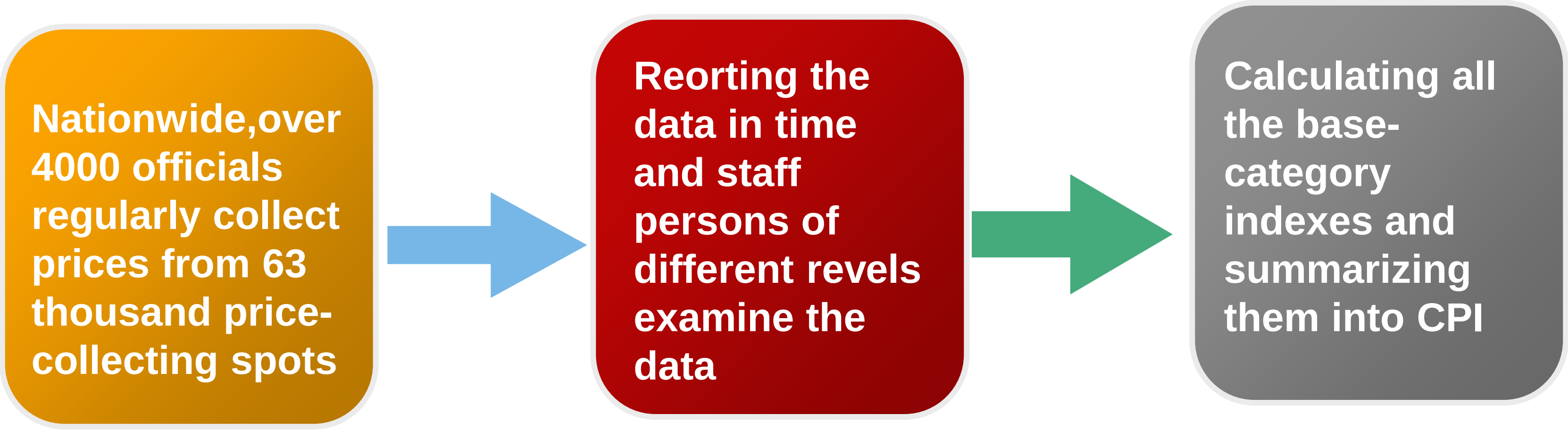
- Necessary
  - In recent years, E-commerce websites such as B2C (www.amazon.com, www.jd.com) and C2C (www.ebay.com, www.taobao.com) development is very rapidly,greatly influencing the individual consuming behavior.
  - The big data in Cyberspace itself suggests the official statistical department apply the data to CPI。
- Difficulties the official statistics department has
  - Compared with E-commerce companies(data owner),the official statistical department (external data achiever) has difficulties in obtaining real data such as transaction price and sales volume.
  - It is difficult to determine the real network sales behavior.
  - It is difficult to determine the sales between e-commerce and traditional retail business.

Background

Current working mode of CPI in official statistics department

Current working mode of CPI in China

- **Data Collection：**Personal Visit。
- **Weight number acquistion：**From Household Income and Survey (133 thousand urban and rural Households in this turn)。
- **Formula：** Chained Laspeyres Formula。
- **Quality adjustment：**The international general method。



Nationwide,over 4000 officials regularly collect prices from 63 thousand price-collecting spots

Reorting the data in time and staff persons of different revels examine the data

Calculating all the base-category indexes and summarizing them into CPI

Outline

1 Backgound

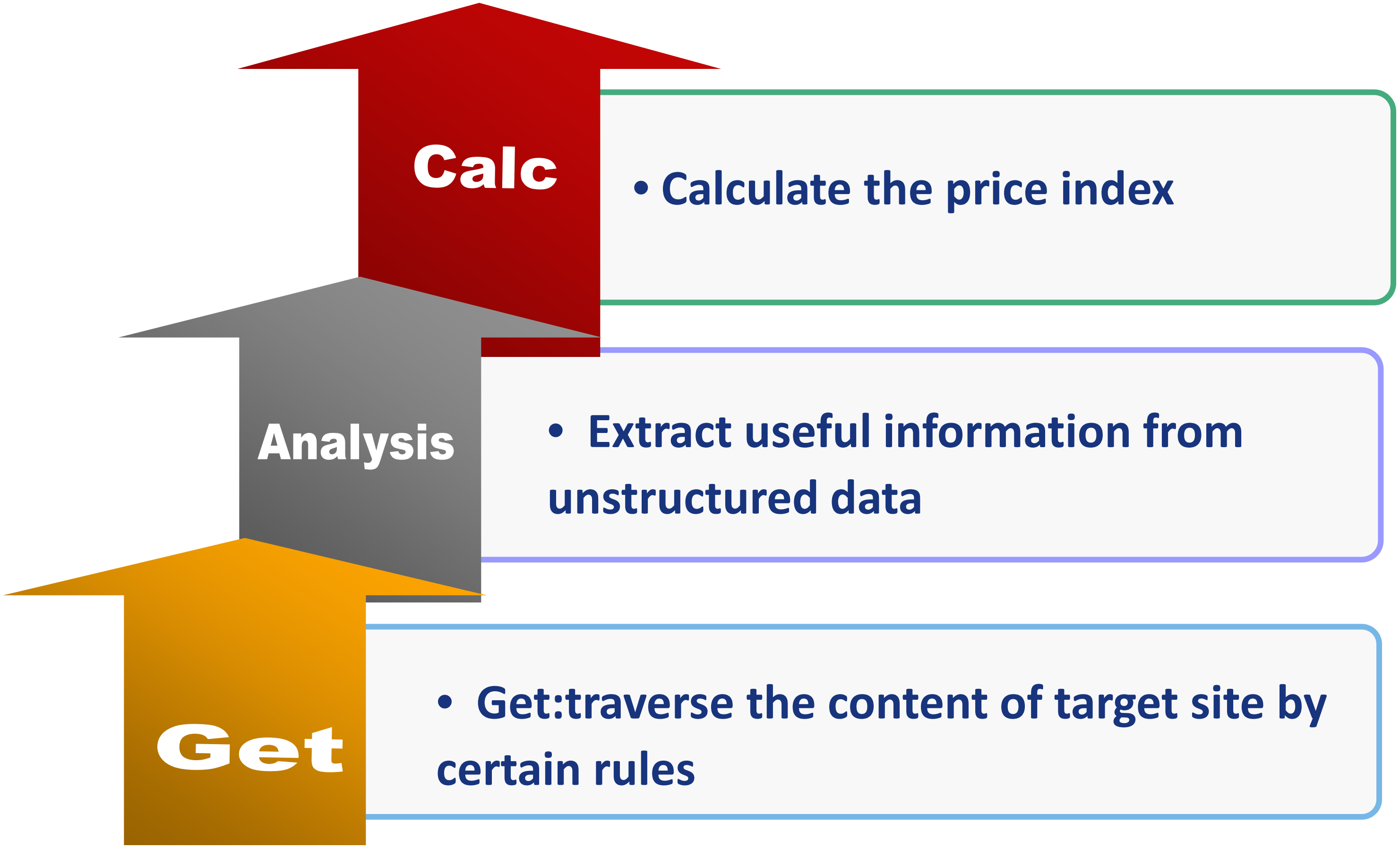
2 Web crawler and the Data characteristics

3 The price index model based on the web crawler and the empirical

4 Advantages and disadvantages

Web crawler and the Data characteristics

External data achiever：work flow based on web crawler



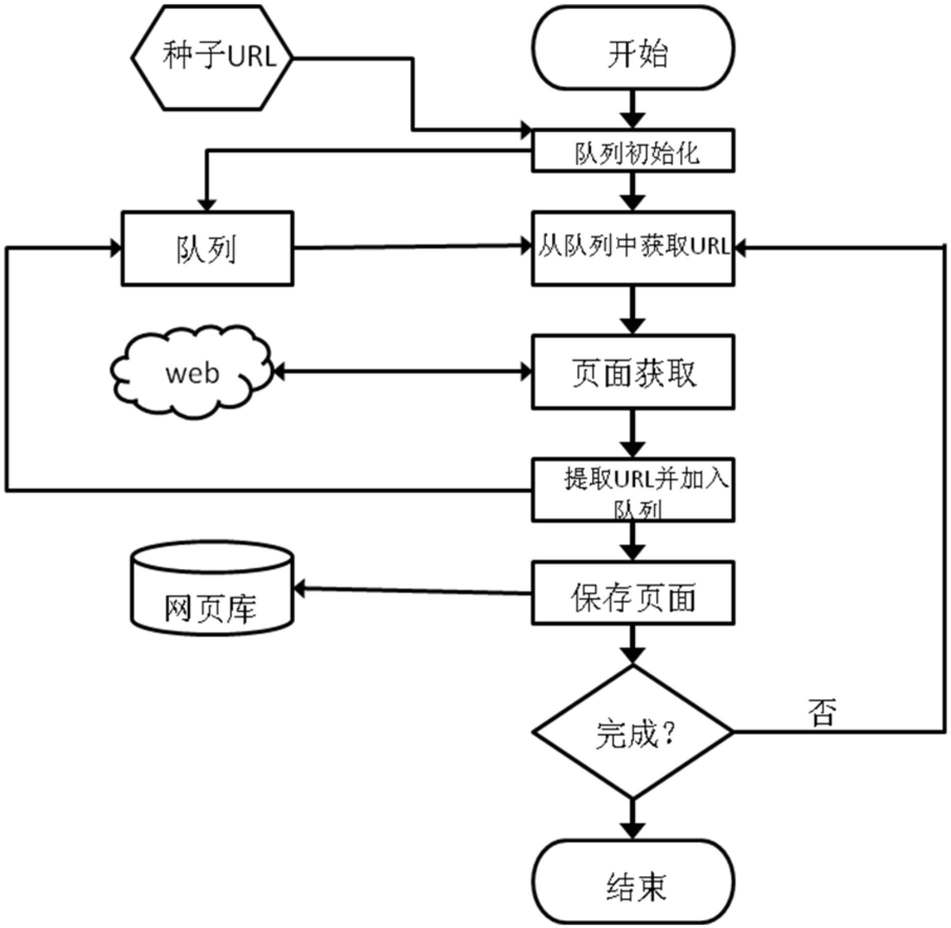
- Calculate the price index

- Extract useful information from unstructured data

- Get:traverse the content of target site by certain rules

Web crawler:traversing the target web site by certain rules, and fetching useful information according to specified format.

Web Crawler



Writing web crawler program used by Perl language in the environment of Linux.

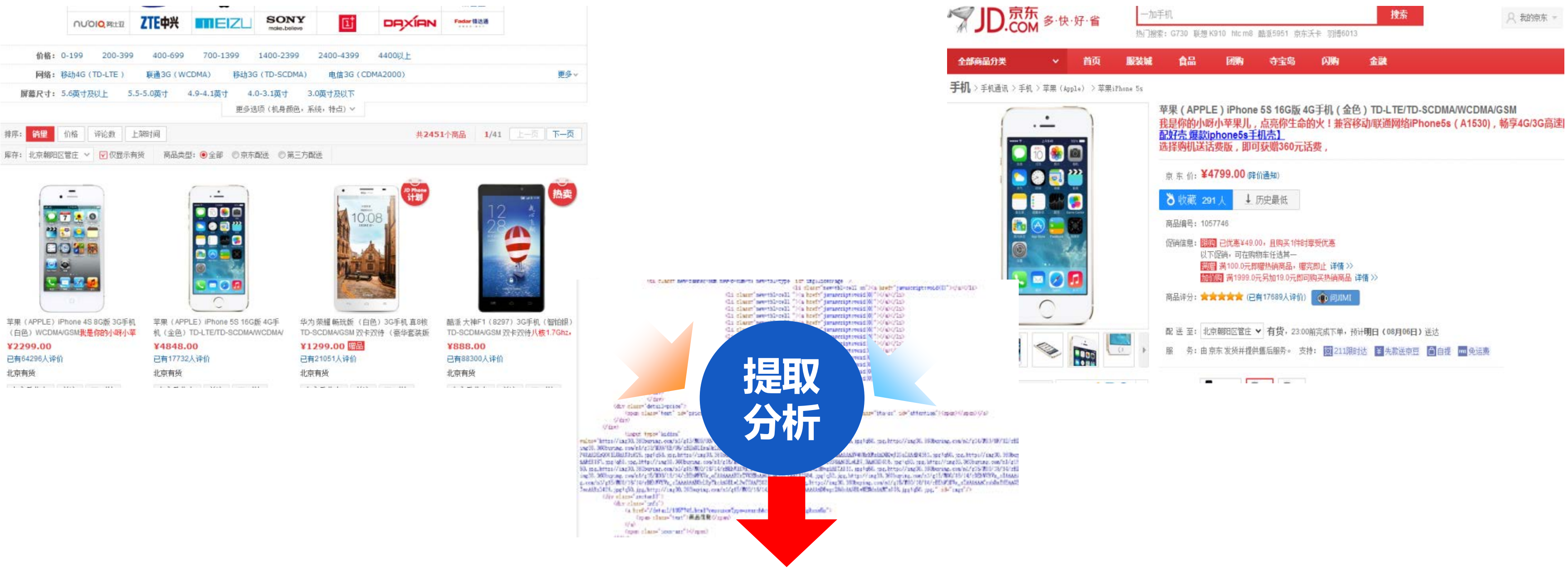
Regular expression could be used to analyze the HTML files which web crawler has fetched.

For example: daily data around 200MB Mobile in phone category of www.JD.com.

Example

www.jd.com **Mobile Phone index Html**

Mobile Phone Html



nuoia 努比亚 ZTE中兴 MEIZU SONY make.believe DAXIAN Fadar 锋达通

价格：0-199 200-399 400-699 700-1399 1400-2399 2400-4399 4400以上

网络：移动4G（TD-LTE ） 联通3G（WCDMA） 移动3G（TD-SCDMA） 电信3G（CDMA2000） 更多

屏幕尺寸：5.6英寸及以上 5.5-5.0英寸 4.9-4.1英寸 4.0-3.1英寸 3.0英寸及以下

更多选项（机身颜色，系统，特点）

排序：销量 价格 评论数 上架时间 共2451个商品 1/41 上一页 下一页

库存：北京朝阳区管庄 ☑仅显示有货 商品类型：●全部 ○京东配送 ○第三方配送

苹果（APPLE）iPhone 4S 8G版 3G手机（白色）WCDMA/GSM我是你的小呀小苹	苹果（APPLE）iPhone 5S 16G版 4G手机（金色）TD-LTE/TD-SCDMA/WCDMA/	华为 荣耀 畅玩版（白色）3G手机 真8核 TD-SCDMA/GSM 双卡双待（豪华套装版	酷派 大神F1（8297）3G手机（智铂银）TD-SCDMA/GSM 双卡双待八核1.7Ghz,
¥2299.00	¥4848.00	¥1299.00 赠品	¥888.00
已有64296人评价	已有17732人评价	已有21051人评价	已有88300人评价
北京有货	北京有货	北京有货	北京有货

JD.COM 京东 多·快·好·省 一加手机 搜索 我的京东

热门搜索：G730 联想 K910 htc m8 酷派5951 京东沃卡 羽博6013

全部商品分类 首页 服装城 食品 团购 夺宝岛 闪购 金融

手机 > 手机通讯 > 手机 > 苹果（Apple） > 苹果iPhone 5s

苹果（APPLE）iPhone 5S 16G版 4G手机（金色）TD-LTE/TD-SCDMA/WCDMA/GSM

我是你的小呀小苹果儿，点亮你生命的火！兼容移动/联通网络iPhone5s（A1530），畅享4G/3G高速

配好壳 爆款iphone5s手机壳】

选择购机送话费版，即可获赠360元话费，

京 东 价：¥4799.00 (降价通知)

收藏 291 人 历史最低

商品编号：1057746

促销信息：限购 已优惠¥49.00，且购买1件时享受优惠

以下促销，可在购物车任选其一

满赠 满100.0元即赠热销商品，赠完即止 详情 >>

加价购 满1999.0元另加19.0元即可购买热销商品 详情 >>

商品评分：★★★★★ (已有17689人评价) 问JIMI

配 送 至：北京朝阳区管庄 有货，23:00前完成下单，预计明日（08月06日）送达

服 务：由 京东 发货并提供售后服务。 支持：211限时达 先款送京豆 自提 免运费

example

	Mobile Phone	Price (yuan)	In stock	Good comment	General comment	Bad comment
1	Samsung B309i	168.00	yes	26939	1415	428
2	iPhone4S 8GB	2448.00	yes	54243	2453	1822
3	Galaxy S4 I9300	1899.00	yes	39131	1873	982

Difference between the two kinds of data

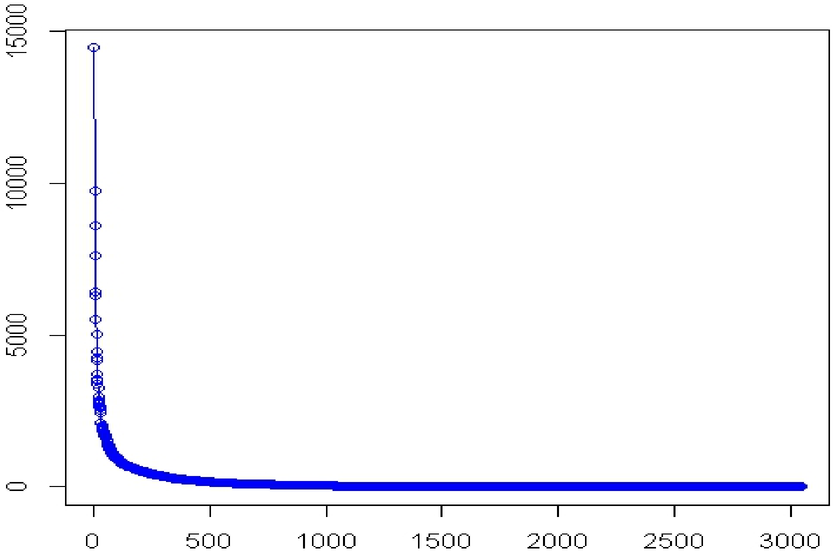
Difference between the two kinds of data

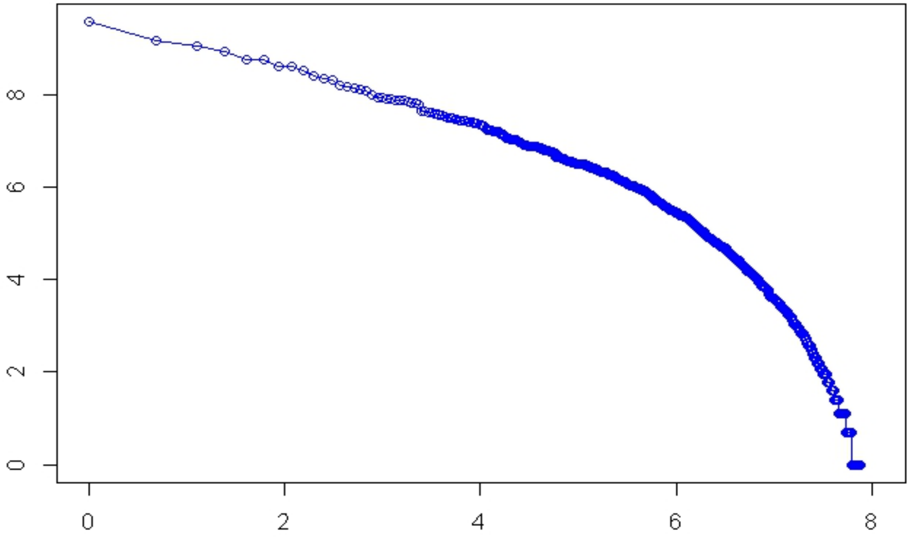
	Web crawler	In field
frequency	1 day	3-5 days
number	Abundant goods number. The types of mobile phone can reach 3000 in one E-commerce website.	5-10 items
Trend of price	volatility	stable
Sale strategy	E-commerce companies often adopt “buy and give” method to promote sales ,difficult to knon real price	Decide whether the price is real or not by examining sales activity in field.

Using cluster analysis to analyze 2 kinds of prices, E-commerce and traditional trail sales have totally different pricing strategy.

How to deal with the challenges brought by “Miaosha” “low price at fixed time”?

Long tail





X:Orderd number order by the volume of comments

Y: the volume of comments

In normal coordinate

X:Orderd number order by the volume of comments

Y: the volume of comments

In double logarithmic coordinate

Data shows that the number of comments of E-commerce is power-law distributed. According to statistics, the sales volume in Amazon is also power-law distributed. Therefore, it is reasonable to think in a particular period of time, the number of comments is similar to the number of sales.

difficulties in web crawler technology

difficulties in web crawler technology

The website E-commerce companies design is very complicated. The useful information is limited. Some websites even design price tag pictures to prevent others from obtaining data

E-commerce companies often redesign their website to attract more customers. But since web crawler technology heavily depends on text parsing strategy, it is necessary to update its strategy according to the change of websites.

Difficulties in fetching data due to website safety strategy.

Outline

1 Backgound

2 Web crawler and the Data characteristics

3 The price index model based on the web crawler and the empirical

4 Advantages and disadvantages

price index (based on web crawler)

In accordance with the current methods, Jevons formula is adopted to calculate the basic category price index(such as mobile phone).

scheme 1：the top 5 mobile phones
scheme 2：all the mobile phones （3000）

3 indexes（last month =100）

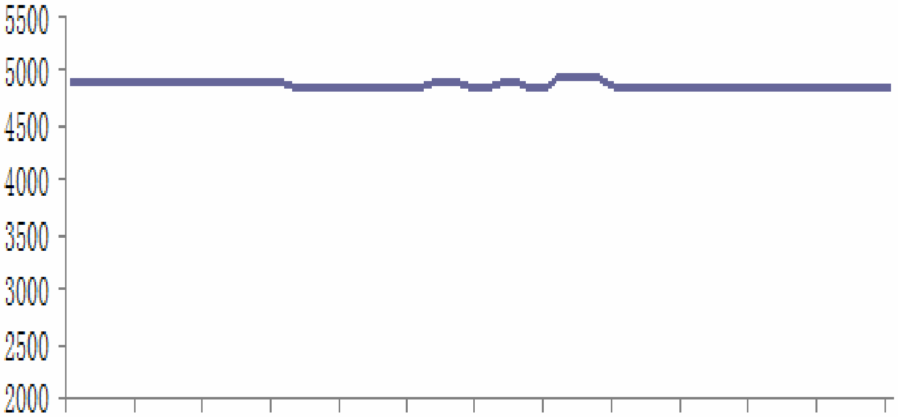
	Index （scheme 1）	Index （scheme 2）	Index （mobile phone index in china cpi）
2014.06	99.0	95.9	99.8
2014.07	99.6	97.7	99.9
2014.08	99.7	96.0	99.6
2014.09	99.5	96.5	99.5

The calculation method based on the web crawler is feasible

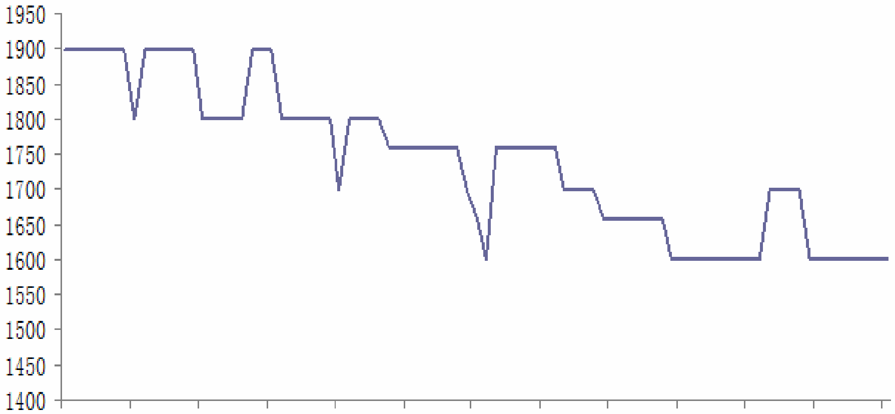
1.The calculation method based on the web crawler is feasible.

2.Why is it that data from scheme 2 is obviously lower than that from scheme 1 and official data?

Take the cell phone for example. The prices of those popular phones like iPhone5S are high and stable for a long time while those small brands try to occupy the market by reducing prices rapidly. Regardless of the equal weight , the price reducing effect is enlarged, resulting in data distortion.



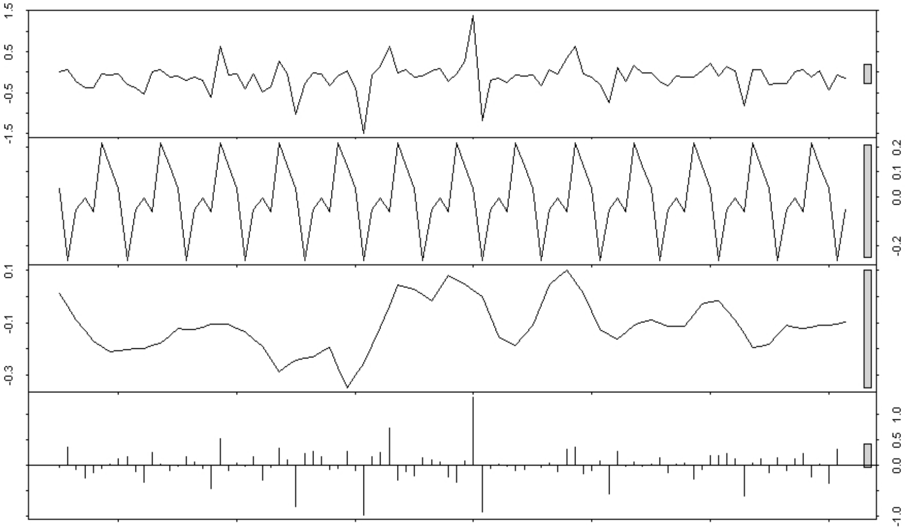
2014.05-2014.09 iPhone 5S



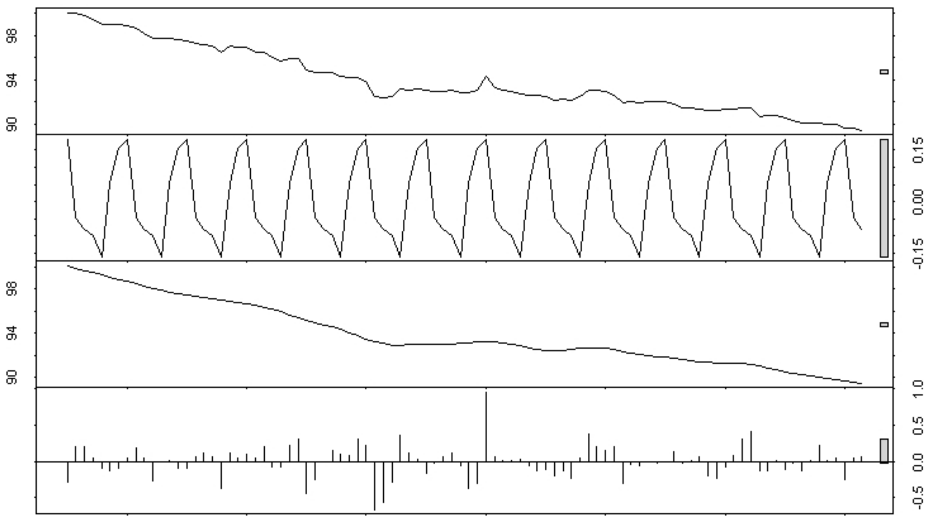
2014.05-2014.09 HTC D816W

Price fluctuation can be observed within shorter period of time

Seasonal adjustment（one week） used by STL method



Index data （last day =100） in scheme 2
Index data(last day =100) = seasonal data + trend data + remainder data



Fixed base price number(2014.05.11=100) in scheme 2
Fixed base price number = seansonal data + trend data + remainder data

1. Similar to the " weekend effect" in traditional retailing, the targeted E-commerce companies have regular sales activity in a week, which is also in accordance with the original data analyzing.

2. The fact that cell phone prices are going down is obviously observed through fixed base index of price in scheme 1 and scheme Two.

Outline

1 Backgound

2 Web crawler and the Data characteristics

3 The price index model based on the web crawler and the empirical

4 Advantages and disadvantages

Advantages

- Official statistical department could widen price-collecting methods through web crawler technology (no legal issues).

- Compared with price-collecting in field, web crawler technology saves time and labor, achieving all-sided statistics.

- Price fluctuation can be observed within shorter period of time

disadvantages

- **How to distinguish real transaction ?**

  - Sales online may involve returns of goods and the real sales volume could be negative. Sometime, it is not representative in statistical sense

  - **How to deal with the problem of representativeness?**

Less is More!The current basic class-index-number calculating method is equal weight . Take all the goods into calculating, weight must be considered otherwise representativeness is lost.

Thank You！