Uses of Network Sentiment Index for Forecasting

Guihuan Zheng

May.2015

利用网络舆情指数进行预测

郑桂环

2015年5月

Contents

- Network big data and sentiment index
- What the network sentiment index can do
- How to build the sentiment index
- Practical cases
- Conclusions and outlook

目录

* 网络大数据与舆情指数
* 网络舆情指数能做什么
* 如何构建舆情指数
* 实际应用案例
* 结论与展望

1. Network big data and sentiment index

- The conception and characteristics of big data
- Classifications of network big data
- The applications of network big data on macroeconomics

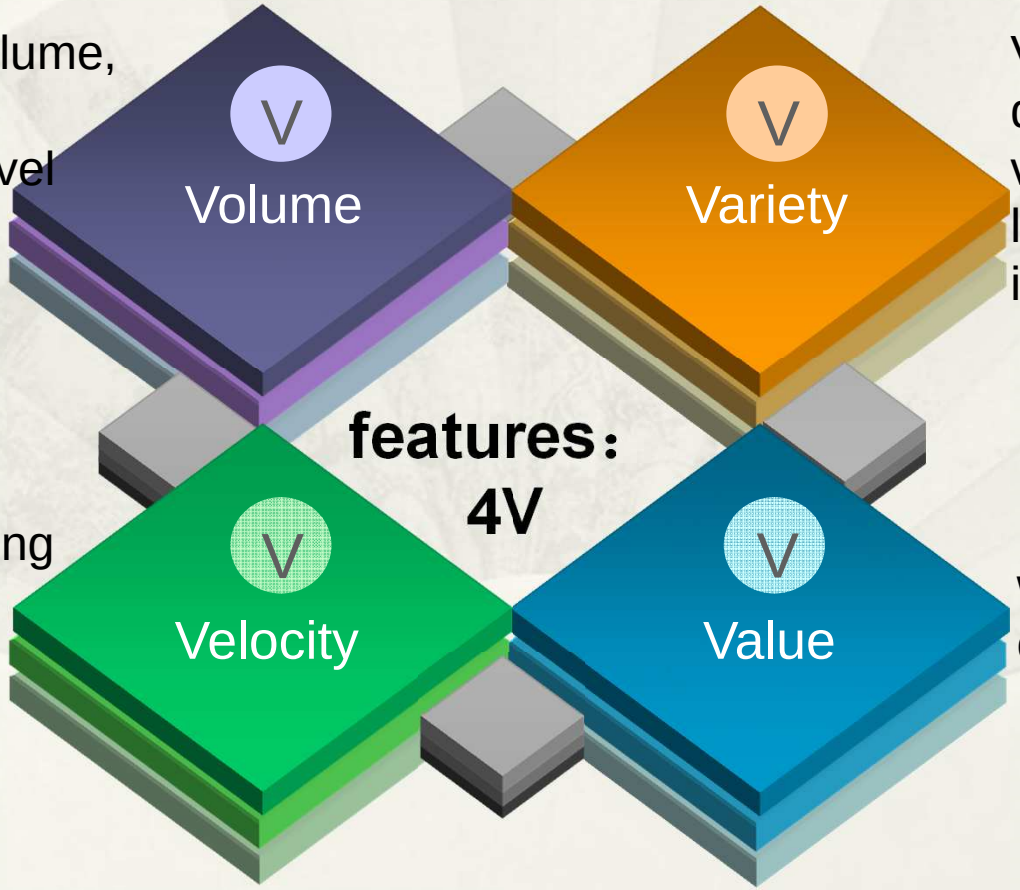
一、网络大数据与舆情指数

* 大数据的概念与特点
* 网络大数据的分类
* 网络大数据在宏观经济领域的应用

1.1 The conception and characteristics of big data

What is big data?

Huge data volume, jump from TB level to PB level



Various kinds of data, like web log, video, picture, location information

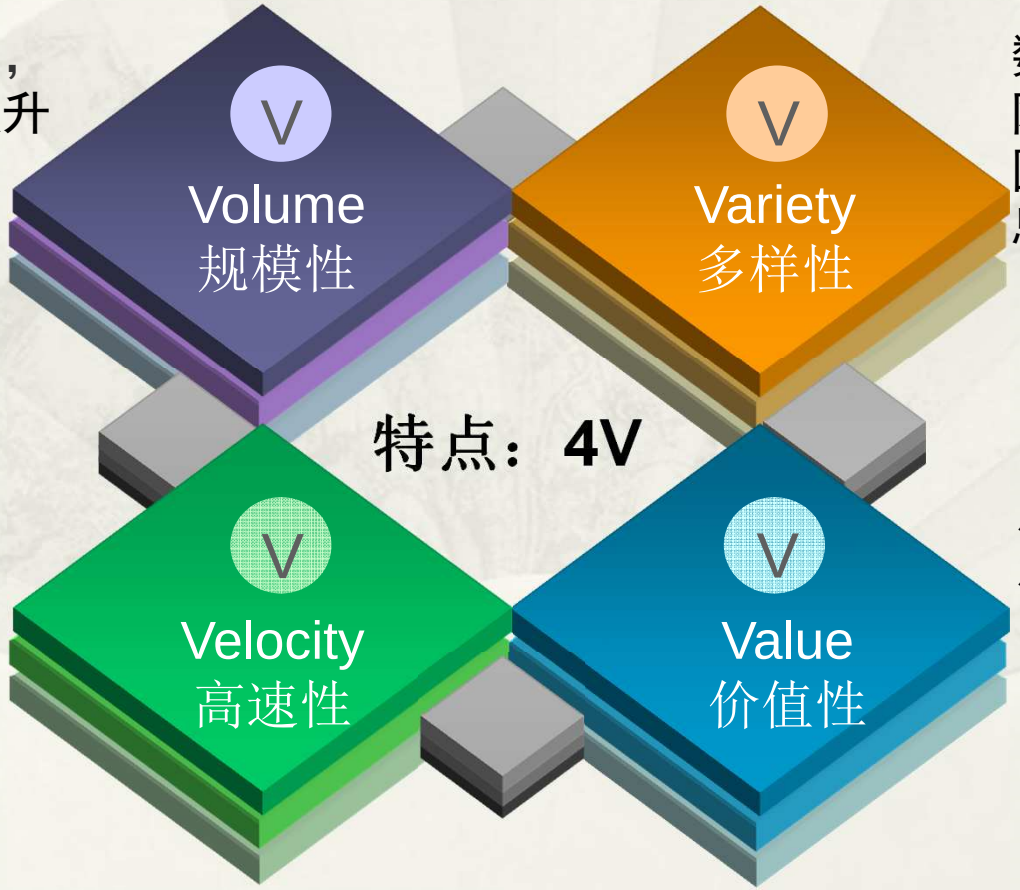
Fast processing speed

Low value density with high commercial value

1.1 大数据的概念和特点

什么是大数据？

数据体量巨大，从TB级别，跃升到PB级别



数据类型繁多。有网络日志、视频、图片、地理位置信息等等

处理速度快

价值密度低，商业价值高

1.2 Classifications of network big data

- Network big data includes news data, browsing data, searching data and interactive data .
  - News data: network news data
  - Browsing data : reflects user's footsteps online , mainly used for consumer behavior analysis by online retailers.
  - Searching data : mainly refers to the time-series searching frequency data of keywords which reflects user's interest, focus and intention.
  - Interactive data : mainly refers to data from micro-blog, Wechat and SNS, reflect the user's preferences and mood 。

1.2 网络大数据的分类

- 网络大数据包括新闻数据、浏览数据、搜索数据、互动数据。
  - 新闻数据
  - 浏览数据：主要用于电商领域的消费者行为分析，浏览数据反映了用户每一步的访问脚步
  - 搜索数据：主要指搜索引擎记录的关键词被搜索频次的时间序列数据，能反映数亿用户的兴趣、关注点、意图
  - 互动数据：主要是微博、微信、社交网站的数据，反映用户的倾向性和情绪因素。

1.3 The applications of network big data on macroeconomics

- Shortages of the traditional survey and Statistics : mainly use the statistics and survey data on Macroeconomic analysis and forecasting, which cost much and exist time lag
- Advantages of network big data： timeliness with low cost for network big data, which makes up for the shortages of traditional survey and Statistics data
- Advantages of network sentiment index based on big data: building network sentiment index with useful information extracting from huge and complex network big data , has the advantages of real-time and intelligence.

1.3 网络大数据在宏观经济领域的应用

- 传统调查统计的不足：在宏观经济监测预警与预测分析中，主要使用统计数据与调查数据，但调查统计的成本较高，同时数据发布存在一定的滞后性。
- 网络大数据的优势：网络大数据的获取成本低，在及时性方面具有优势，可以弥补统计与调查抽样的不足。
- 基于网络大数据构建网络舆情指数的优势：从庞大而复杂的网络大数据中提取有用信息，构建网络舆情指数，具有实时性和智能化的优势 。

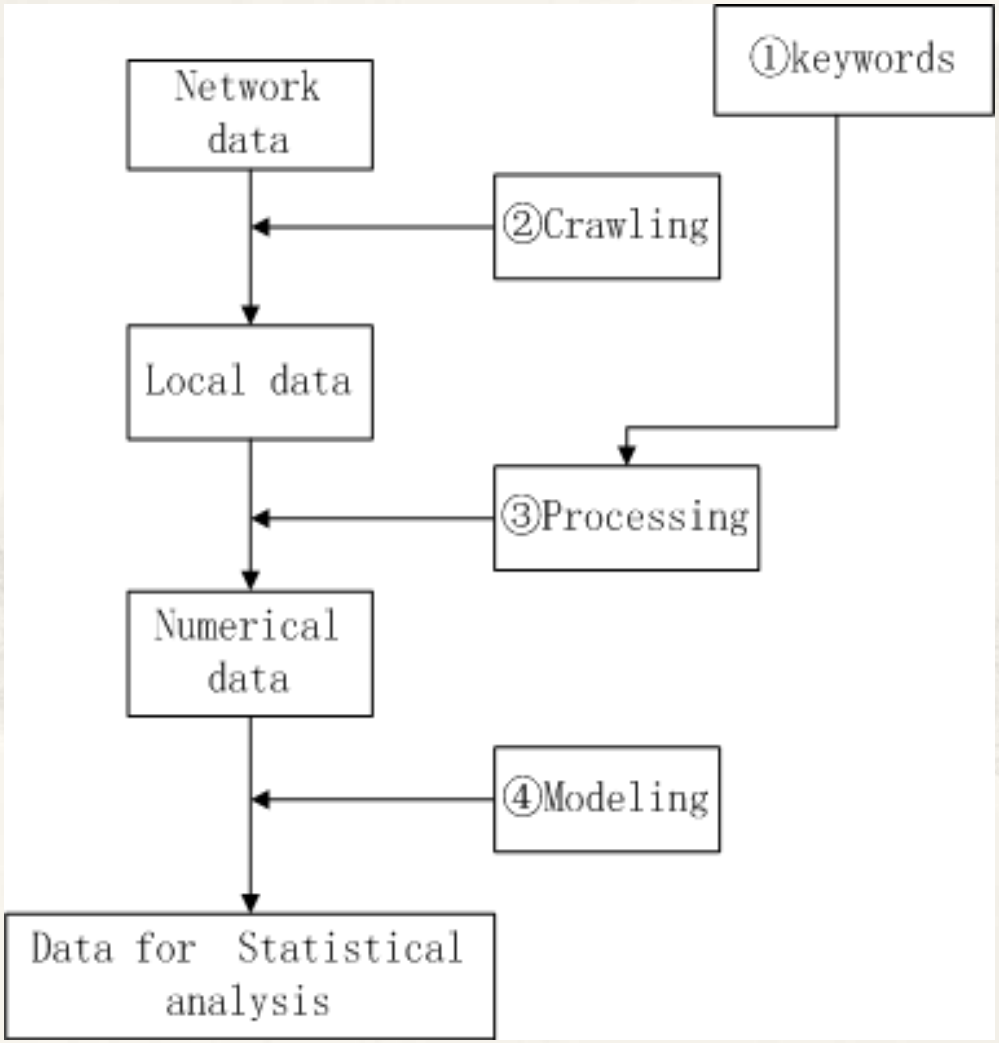
2. What the network sentiment index can do

- **Can make the unmeasurable expected variable in traditional area become possible, like people's expectation for stock market.**
  - By screening and integrating network data, we can mine the sentiment of economic market and build the sentiment index.
- **Construction of real-time monitoring index for high-frequency monitoring and real-time forecasting**
  - E.g., By using daily network news data to build inflation sentiment index, we can now-cast current month CPI and achieve real-time forecasting.
- **Build main macroeconomic leading index for early warning and forecasting**
  - Compared with traditional leading index, the index based on network big data can reflect the subjective feelings and expectation ,may become a new statistical monitoring method

二、网络舆情指数能做什么

- 使许多传统领域不能预测度量的预期变量变为可能。例如社会对股市的预期。
  - 通过对网络数据的筛选和融合，可以做到对经济市场的情感挖掘，构建预期指数
- 构建实时监测指数用于高频监测和实时预测
  - 例如，使用网络日度新闻数据构建通胀舆情指数，预测当月CPI，实现实时预测
- 构造主要宏观经济领先指数用于预警与预测
  - 与基于统计数据构建的先行合成指数相比，基于网络数据构建的宏观经济领先指数综合的是网民的主观感受和预期判断，是新的统计监测渠道。

3. How to build the sentiment index

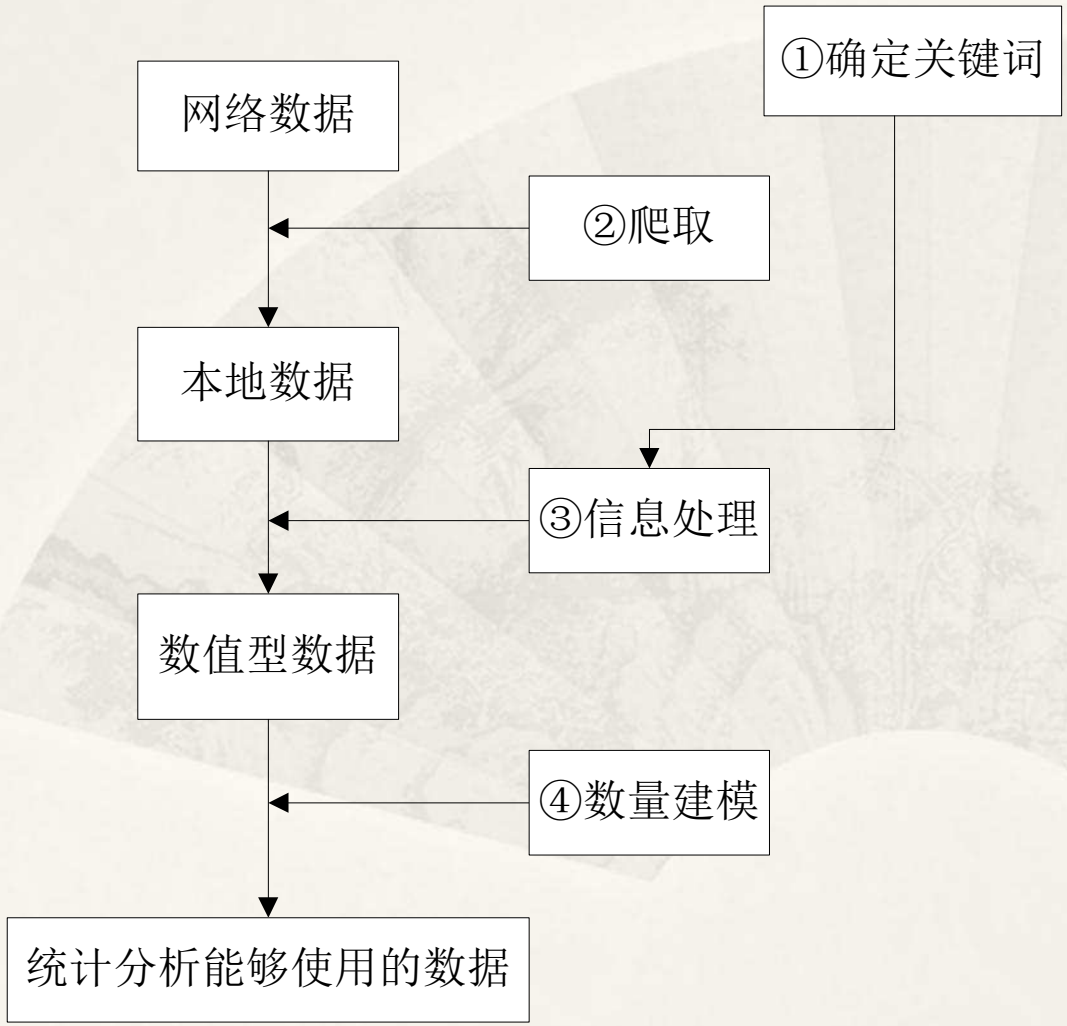


Keywords, e.g. commodity price

Rose category: prices rise, price rally, price increase, price boost, price upward;

Drop Category: price fall, price drop, price decrease, price decline、price downward

三、如何构建网络舆情指数



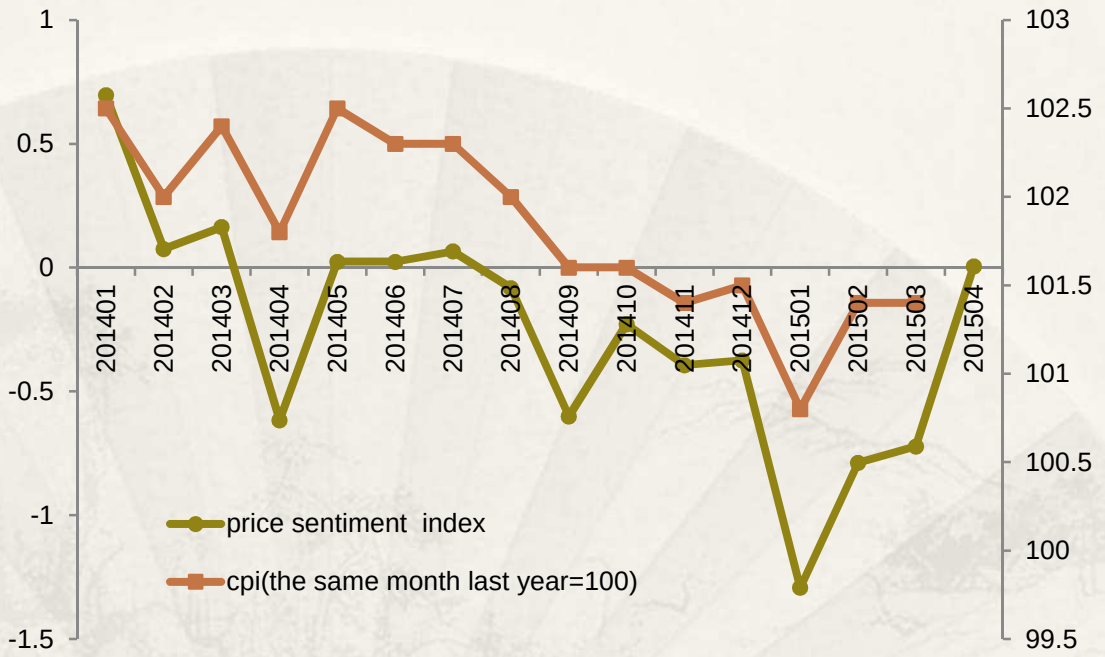
关键词确定，以物价为例。

上升类：物价上升、物价回升、物价上涨、物价上行；

下降类：物价下滑、物价下跌、物价回落、物价下行、物价下降

4、Application 1: build price sentiment index for now-casting

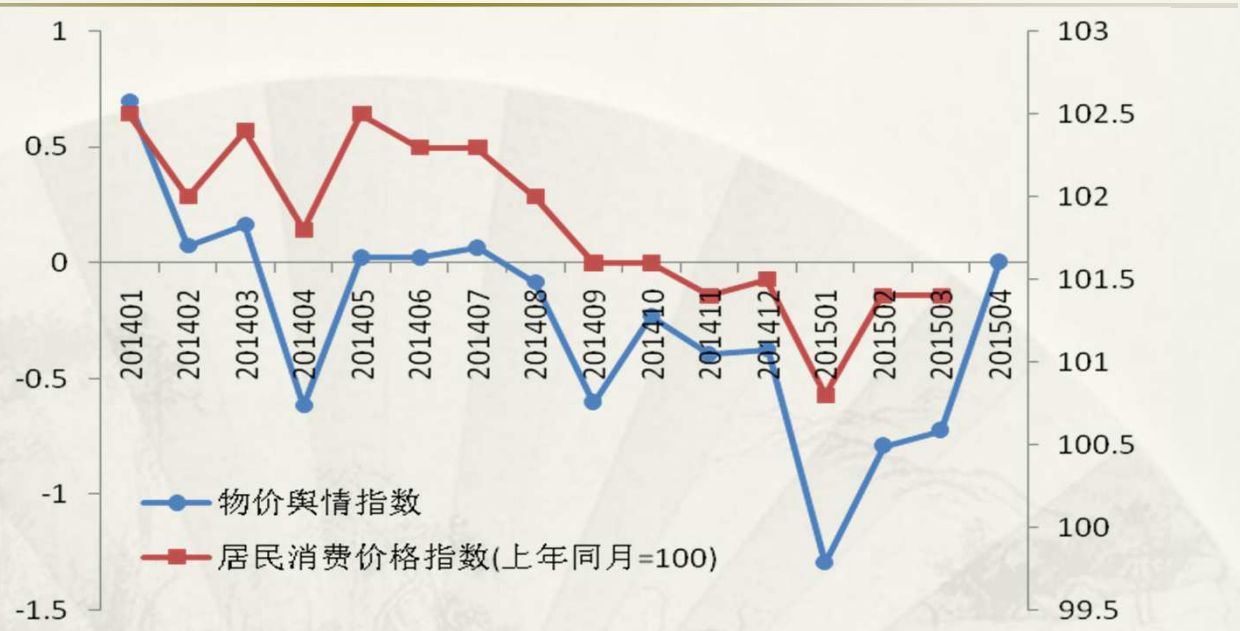
date	basic data	
	news	
	up classs	down class
201301	3,425	634
201302	5,001	253
201303	4,617	359
201304	3,247	703
201305	3,494	606
201306	2,165	262
201307	3,672	478
201308	2,672	477
201309	3,822	423
201310	2,973	506
201311	1,726	717
201312	4,083	421
201401	3,781	874
201402	2,836	348
201403	3,987	514
201404	2,576	697
201405	2,877	510
201406	2,850	495
201407	1,534	523
201408	2,124	603
201409	3,809	701
201410	2,178	700
201411	2,192	1,045
201412	4,877	1,154
201501	5,329	1,973
201502	1,390	1,492
201503	4,021	1,236
201504	1,011	988



1. Use network news data
2. Historical trend : volatility is very consistent . Can be used to determine trends and turning points.
3. High-frequency price sentiment index based on daily data before April.20,2015, indicates April's CPI will rise significantly

四、实际应用案例之一：构建物价舆情指数用于实时预测

日期	基础数据	
	新闻	
	上行类	下行类
201301	3,425	634
201302	5,001	253
201303	4,617	359
201304	3,247	703
201305	3,494	606
201306	2,165	262
201307	3,672	478
201308	2,672	477
201309	3,822	423
201310	2,973	506
201311	1,726	717
201312	4,083	421
201401	3,781	874
201402	2,836	348
201403	3,987	514
201404	2,576	697
201405	2,877	510
201406	2,850	495
201407	1,534	523
201408	2,124	603
201409	3,809	701
201410	2,178	700
201411	2,192	1,045
201412	4,877	1,154
201501	5,329	1,973
201502	1,390	1,492
201503	4,021	1,236
201504	1,011	988



1、使用网络新闻数据。

2、历史走势：波动较为一致。可用于趋势和拐点判断。

3、根据截止2015年4月20日的日度数据编制的物价高频舆情指数显示，4月份CPI同比将明显回升。

4. Application 2: build unemployment sentiment index for early warning and prediction

* Based on Baidu searching data

keywords	Selected reasons	weight
Unemployment card Unemployment insurance Unemployment benefits	Directly reflect the dynamic changes of the unemployment Generally, only the unemployed will care for unemployment insurance Acts as Unemployment insurance	Every indicator: 15%
Laid-off venture Reemployed Laid-off card loans	Reflect the changing number of laid-off venture for the unemployed. Reflect the changing number of reemployed Reflect the changing number of laid-off venture for the unemployed	Every indicator: 10%
Layoffs Science and Technology Recruitment Financial crisis inflation	Indicate corporate layoffs trend, which affects the number of unemployed people Indicate scientific development, which affects the number of unemployed people Reflects the changes in the number of the unemployed Financial crisis affect the number of the unemployed indicate current inflation, which affects the number of the unemployed	Every indicator: 5%

四、实际应用案例之二：构建失业率舆情指数用于预警预测

* 基于百度搜索引擎数据

关键词	入选原因	权重
失业证	直接反映失业人群的动态变化	每个指标15%
失业保险金	一般情况下只有失业者会关心失业保险金	
失业金	作用同失业保险金	
下岗创业	反映一部分失业人群选择下岗再创业的数量变化	每个指标10%
下岗再就业	反映一部分了解失业再就业情况的失业人群的数量变化	
下岗证贷款	反映一部分失业人群选择下岗再创业的数量变化	
裁员	反应企业裁员趋势，从而影响失业人群数量的变化	每个指标5%
科学技术	反应科学发展情况，从而影响失业人群数量的变化	
招聘	反应当前失业人群数量的变化	
金融危机	反应当前是否发生金融危机，从而影响失业人群数量的变化	
通货膨胀	反应当前通货膨胀的程度，从而影响失业人群数量的变化	

Data processing

- standardization

$$\text{Current search volume} = \frac{\text{original volume of current search}}{\text{original volume of initial search}}$$

- synthesis

$$W = \sum(\text{key} \times \text{weight})$$

key represents the search volume, weight is the corresponding weights

- Apply moving average smoothing method to deal with the series

数据处理

* 标准化

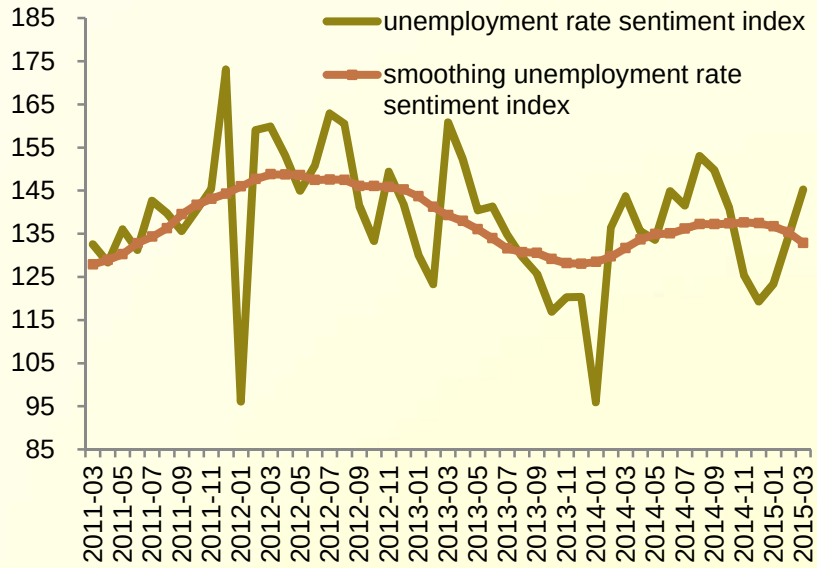
$$\text{当期搜索量} = \frac{\text{原当期搜索量}}{\text{原初期搜索量}} \times 100$$

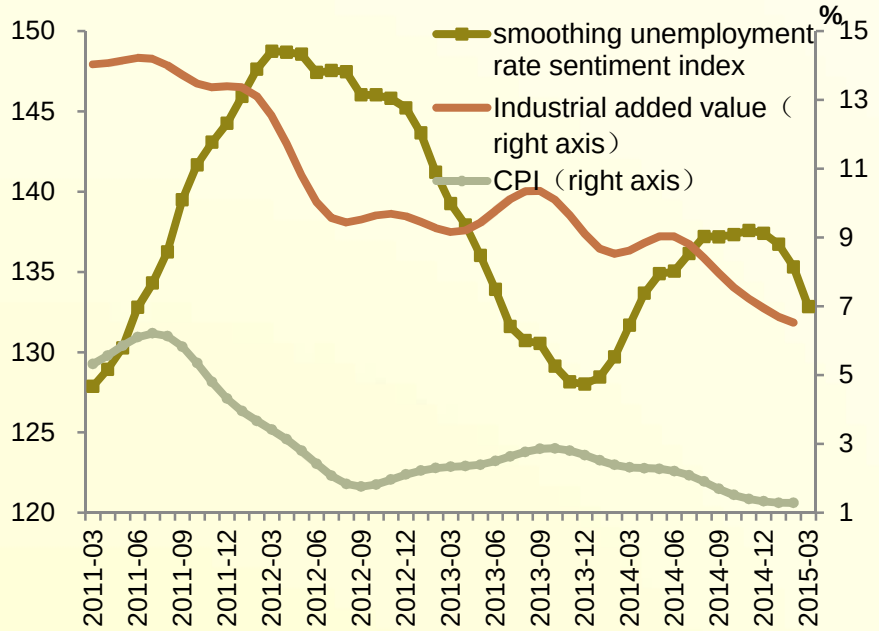
* 合成

$$W = \sum(key \times weight)$$

其中，key表示每个关键词的搜索量，weight表示对应的权重。

* 对合成后的序列使用移动平均进行平滑处理

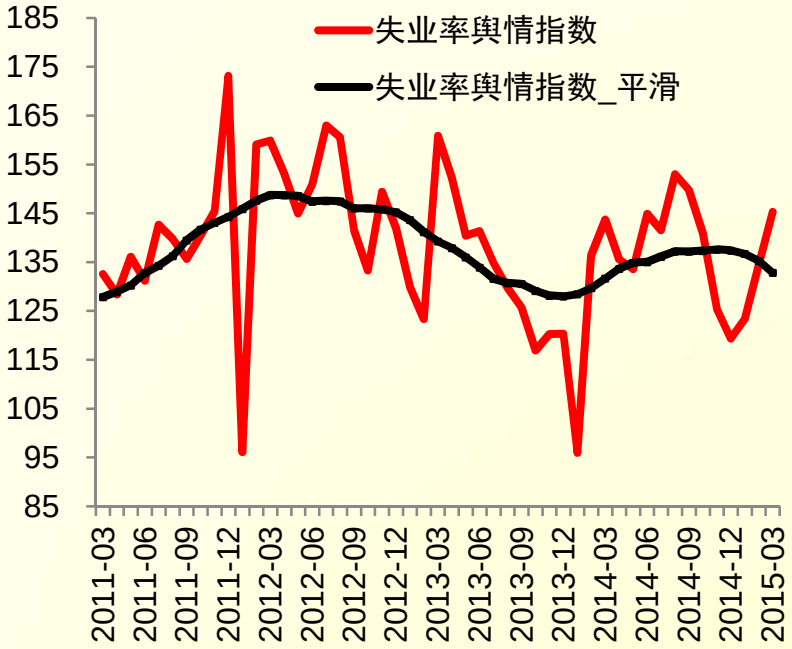


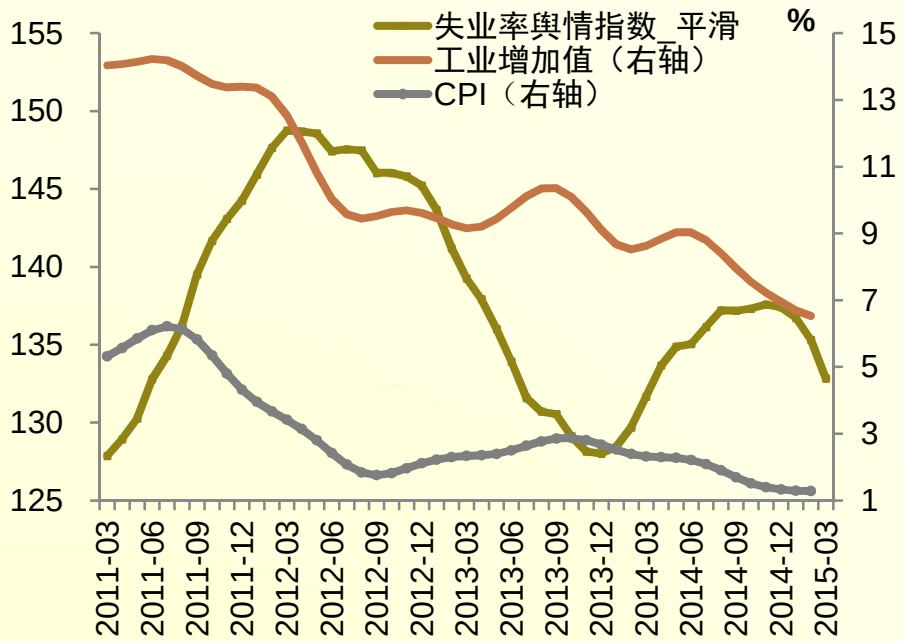


1. The index is negatively correlated with industrial added value and CPI, which is consistent with Okun‘s law and Phillips curve . The index is leading indicators. Leading industrial added value 9-12 months, leading CPI 7-9 months.

2. It can provide warning signals in advance.

（1） In early 2011, the index has been rising fast . Industrial added value and CPI began to decline in the second half of 2011. It sent a signal for the economic downturn.（2） The index began to rise significantly in the fourth quarter of 2013, then the Industrial production and CPI showed a downward trend in the second quarter of 2014, It also sent a warning signal for the economic downturn.



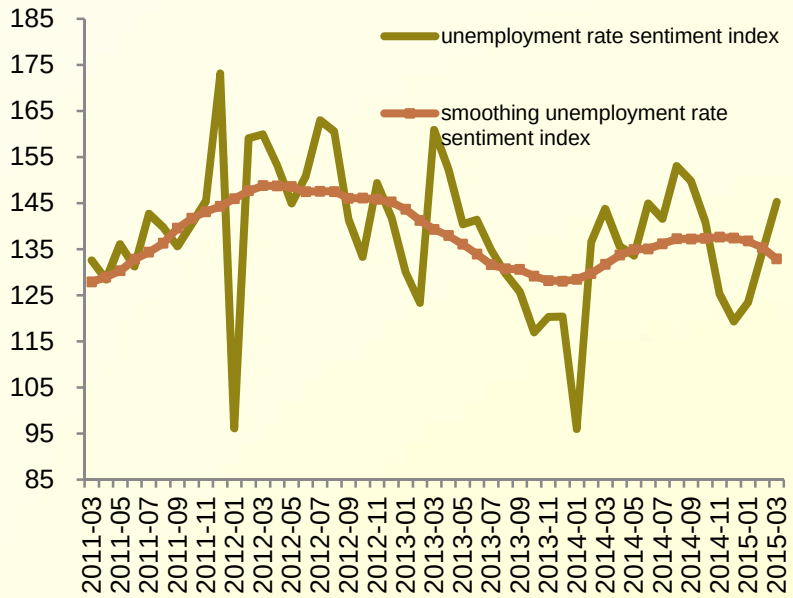


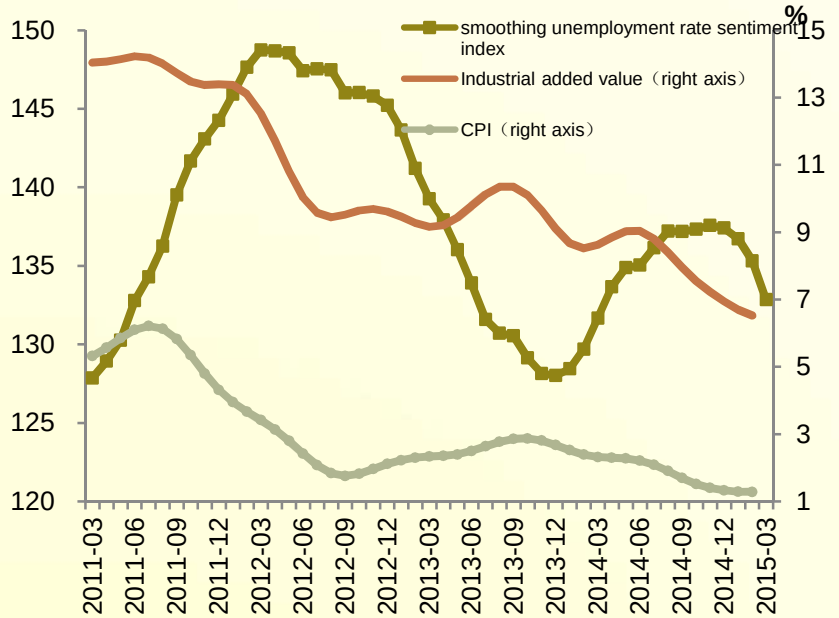
1、失业率舆情指数与工业增加值、CPI呈负相关关系，符合奥肯定律和菲利普斯曲线，并且呈现一定的先行性。领先工业增加值9-12个月；领先CPI7-9个月。

2、可以提前发出预警信号。

（1）2011年初失业率舆情指数一直快速上行。工业增加值和CPI于2011年下半年开始下滑。失业率舆情指数为经济下行发出了预警信号。

（2）2013年第四季度失业率舆情指数开始显著上扬，工业增加值和CPI于2014年第二季度开始呈下行态势，失业率舆情指数为这次的经济下行也发出了预警信号。



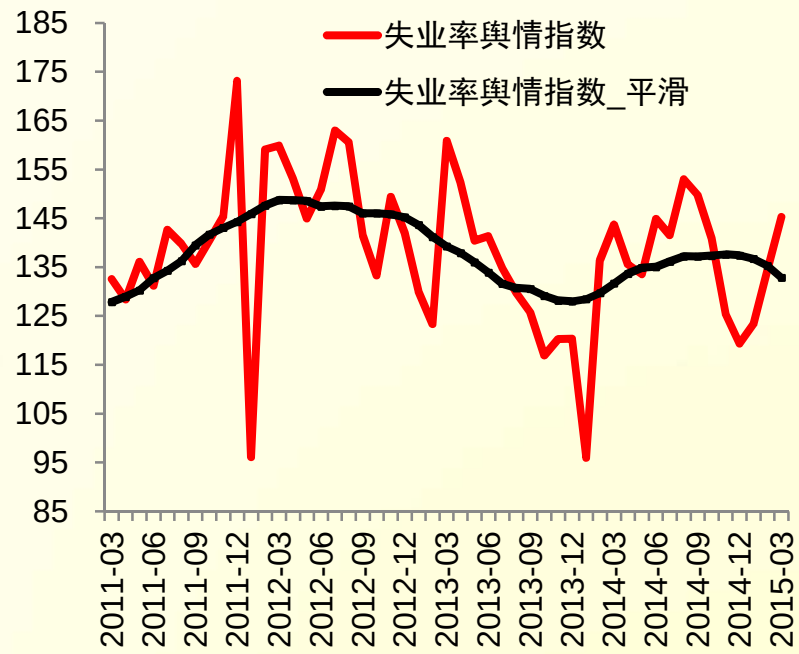


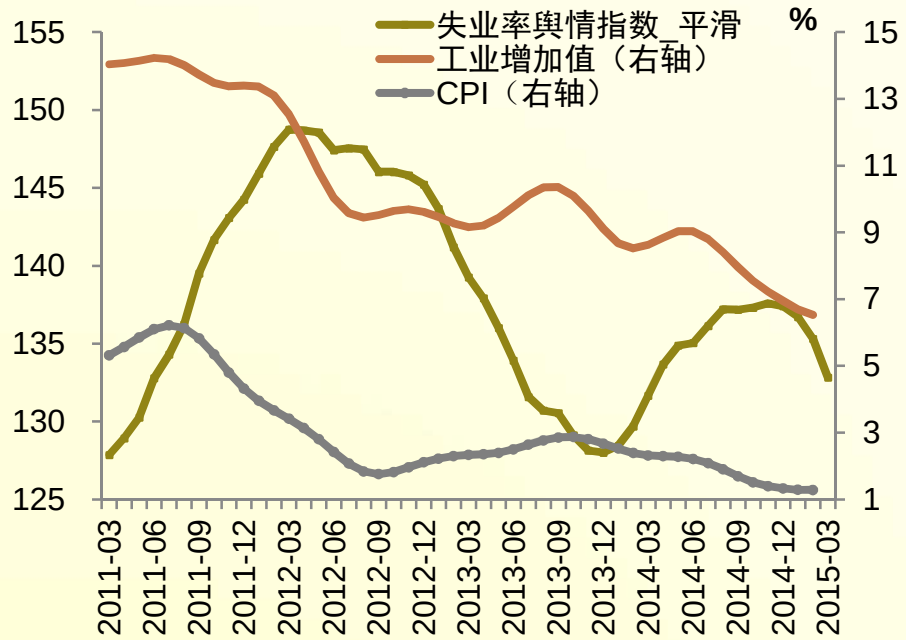
The current trend shows: economic fundamentals is under control, but need policy support

（1） The current upward trend slows down, showing economic fundamentals is controlled, there is no further deterioration in the job market.

（2） According to the rapid rise of the index, there is a downward pressure on the economy in the first half of this year.

（3）the recent fell of the index has a great relationship with the Chinese New Year, the original data shows the March index has significantly risen. So the trend is not stable, with the possibility of rebound. Need policy support to prevent the deterioration of economic fundamentals.





当前走势显示：基本面可控，但还需要相关政策继续给予支持。

（1）当前上扬态势明显趋缓，显示基本面可控，就业市场压力没有进一步恶化。

（2）根据舆情指数前期快速上扬的态势判断，今年上半年经济下行压力较大。

（3）近期舆情指数下降与1、2月份受春节效应影响有很大关系，原始数据的走势显示，3月份失业率舆情指数已经显著走高。因此，失业率舆情指数放缓的走势并不稳固，存在反弹的可能性，还需要相关宏观政策继续给予一定的支撑，以防止经济基本面发生恶化。

4. Conclusions and outlook

- In the era of big data, network big data provides a new data source.
- How to make effectively use of big data as a complement for the latest statistic, it is a new question.
- How to effectively use the network big data to improve the macroeconomic early warning, analysis and trends prediction, it is a new challenge.

四、结论与展望

* 大数据时代，网络大数据提供了新的数据来源。
* 如何有效利用网络大数据作为现有短期统计的补充？是一个新的问题。
* 如何有效利用网络大数据完善宏观经济的预警模型，对宏观经济的走势进行预测，改善经济分析和预测工作，是一个新的挑战。

Thank You !

Thank You !