网络文本挖掘方法介绍

——网络文本挖掘研究系列专题之一

夏潇阳 分析师
电话：021-60750625
eMail：xxy2@gf.com.cn
执业编号：S0260512030005

文本信息的重要性

现代金融分析中，数据无疑具有极端重要的位置，但是随着研究需求的提高和深入，对于数据的要求越来越高，而更多的“非数据信息”也逐渐体现其重要性，比如宏观经济数据一致预期、公告信息挖掘和研究报告挖掘。

网络文本挖掘及其方法

文本挖掘是以半结构（网页）或者无结构（如纯文本）的自然语言文本为对象的数据挖掘，因此又被称为文本数据挖掘。网络文本挖掘的主要步骤如下：首先，从网络上下载文本，我们分析的主要文本包括纯文本、PDF 文本和 HTML 网页等。其次，对于 PDF 文本，我们需要将其转换为纯文本；而对于 HTML 网页，我们需要提取其中的正文信息和其它有用信息（如标题、网址和日期等）。第三，我们对正文信息进行文本“脱水”处理，所谓“脱水”，是指去除文本中的无用信息。第四，我们对脱水后的文本进行模式识别，找出我们需要的信息。最后，我们用金融工程方法对提取出来的信息进行量化分析。

文本下载和文本提取

文本下载可分为文件下载和网页下载两类，文件下载较为简单，Windows 用户可使用开源的 wget 工具进行下载，可用于下载纯文本和 PDF 格式文件。网页下载又分为静态网页下载和动态网页下载两种，静态网页用查看源文件方法看到的源代码与用户看到的最终界面是一致的，而动态网页用查看源文件方法看到的源代码和用户看到的最终界面是不一致的。

读取了 HTML 网页源代码后，我们需要提取其中的正文信息和其它有用信息：首先缩小范围，找到有用信息所在区域，接着找到标题、网址和日期等信息所在位置的特征，然后记录这些信息，最后找到正文以及相应段落所在位置的特征，并提取正文。

提取 PDF 文本时，我们使用 wget 下载 PDF 文件，然后使用工具将 PDF 文件转换为 TXT 文件，接着对 TXT 文件进行细节上的处理，并判断转换是否成功，最后打开 TXT 文件，读入 MATLAB，并输出。

文本脱水和模式识别

将 PDF 转换为纯文本的时候，无法提取 PDF 文件中的格式和排版，因此，提取后的文本不可避免的存在一部分无用信息。文本脱水是指去除文本中的无用信息，文本脱水步骤中的细节很多，其主要想法是，首先尽可能不删除有用信息，然后尽可能删除无用信息。

模式识别和文本脱水所用的方法类似，都需要对文本进行模式匹配，区别在于，文本脱水是删除符合特定模式的文本，而模式识别是找出符合特定模式的文本。

目录索引

图表索引

一、文本信息的重要性

现代金融分析中，数据无疑具有极端重要的位置，但是随着研究需求的提高和深入，对于数据的要求越来越高，而更多的“非数据信息”也逐渐体现其重要性。试举几例：

宏观经济数据一致预期：最近几年，随着几家数据供应商的努力，现在行业公司的一致预期已经相当完善了，但是宏观经济数据的一致预期却一直缺失。如果一个宏观策略研究员想知道目前市场对于CPI、PPI、M1和PMI等数据的市场预期，只能采取“电话调研法”，或者是阅读至少5篇以上近期的相关报告。毋庸置疑，这样的样本显然是有偏的。而对于量化研究员来说，想藉此建立模型也很难做到。

公告信息挖掘：上市公司定期报告中的财务数据是各种分析的基础，但是除此之外，报表附录中的各种信息，非定期的各种公告中也蕴含这很大的信息量。我们去年在“异常交易公告”（未来三个月内不存在重大事项）中找到了超额收益，但这仅仅是开始，形形色色的公告可以带来很多辅助判断的信息。而这些信息如果用人工来分析显然工作量太大，且回溯困难，因此，如果可以用某些工具提取出其中的关键信息，对于这类信息的挖掘是很有帮助的。

研究报告挖掘：随着对于行业研究员一致预期的深入研究，单纯的盈利预测和投资评级信息已经远远不够了。研究员各种欲语还休，各种言外之意，成为买方与卖方之间默契的共识。而另外一方面，研究员的行文本身也是其情绪的一种反映，而这种情绪无疑比公众的情绪更有信息含量。

以上列举的仅仅是文本挖掘在金融领域的一小部分应用，如果没有技术，再怎么好的想法也出不了结果。因此，我们首先介绍如何提取和转化这些文本类的信息。

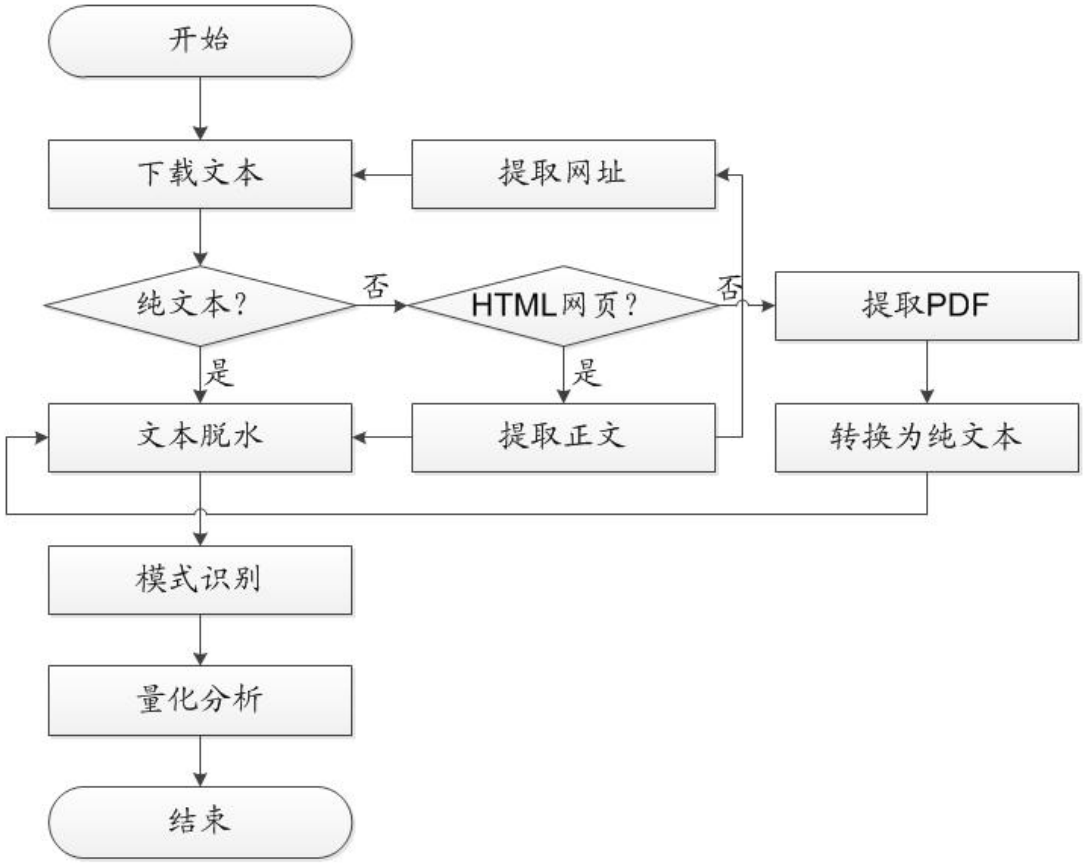
二、网络文本挖掘及其方法

文本挖掘是以半结构（网页）或者无结构（如纯文本）的自然语言文本为对象的数据挖掘，因此又被称为文本数据挖掘。文本挖掘是自然语言处理（Natural Language Processing，简称NLP）的范畴之一。

网络文本挖掘的主要步骤如下：首先，从网络上下载文本，我们分析的主要文本包括纯文本、PDF文本和HTML网页等。其次，对于PDF文本，我们需要将其转换为纯文本；而对于HTML网页，我们需要提取其中的正文信息和其它有用信息（如标题、网址和日期等）。第三，我们对正文信息进行文本“脱水”处理，所谓“脱水”，是指去除文本中的无用信息。第四，我们对脱水后的文本进行模式识别，找出我们需要的信息。最后，我们用金融工程方法对提取出来的信息进行量化分析。

网络文本挖掘的流程如图1所示：

图1：网络文本挖掘的流程图



数据来源：广发证券发展研究中心

三、文本下载和文本提取

文本下载可分为文件下载和网页下载两类，文件下载较为简单，Windows用户可使用开源的wget工具进行下载，可用于下载纯文本和PDF格式文件。wget本身为命令行形式，可被其它编程语言如MATLAB调用。

网页下载又分为静态网页下载和动态网页下载两种，静态网页用查看源文件方法看到的源代码与用户看到的最终界面是一致的，而动态网页用查看源文件方法看到的源代码和用户看到的最终界面是不一致的。对于静态网页下载，可使用MATLAB的自带函数urlread，但urlread不能用于动态网页的下载。我们以某网页为例来展示静态网页和动态网页的区别。

图2为某动态网页在Internet Explorer下的源代码，图3为该网页用urlread读出来的源代码，和图2基本一样，但是这个源代码和用户看到的最终界面是不一致的。图4是我们读出来的该网页的真实源代码，该源代码和用户看到的最终界面是一致的。

图2：某动态网页在Internet Explorer下的源代码

```
        <td height="25" align="center" class="erji_title">每日结算会员成交持仓排名</td>
      </tr>
      <tr>
        <td align="center"> </td>
      </tr>
      <tr>
        <td align="center"><table width="92%" border="0" cellspacing="0" cellpadding="0">
          <tr>
            <td align="right">
                日期：<SPAN class=search>
<input name="actualDate" id="actualDate" type="text" class="Wdate" onfocus="new WdatePicker(this,'%Y-%M-%D',true,'default')" readonly />

                    <!-- <input name="actualDate" id='actualDate' type="text" class="searchinput" size="15" readonly />
                     <a href="javascript:show_calendar('actualDate')" onMouseclick="MM_openBrWindow('show_calendar','','')" style='text-decoration: none;color: #305C83;' >
                    <img src="../../images/i_date_normal.gif" width="30" height="18" align="absmiddle" border='0' />
</a>
-->
</span><span class="search"> <img src="../../images/chaxun.gif" width="47" height="21" align="absmiddle" onclick='ondate()' style='cursor: hand;' /></span>
</td>

          </tr>
        </table></td>
      </tr>
      <tr>
        <td valign="top" height="390">
          <table width="90%" border="0" align="center" cellpadding="0" cellspacing="0">
            <tr>
              <td>



                    <TABLE cellSpacing=1 cellPadding=0 width="100%" align=center border=0>
                        <TD>
      <!--  日期控件  -->
      |
        <div id="textArea" style="font-size:14px;"></div>
        <div id="textArea1" style="font-size:14px;"></div>
        </TD>
                    </TABLE>
```

数据来源：中国金融期货交易所网站、广发证券发展研究中心

图3：某动态网页用urlread读出来的源代码

```
class="erji_title">每日结算会员成交持仓排名</td> </tr> <tr> <td align="center"> </td> </tr> <tr> <td
align="center"><table width="92%" border="0" cellspacing="0" cellpadding="0"> <tr> <td align="right"> 日期：
<SPAN class=search> <input name="actualDate" id="actualDate" type="text" class="Wdate" onfocus="new
WdatePicker(this,'%Y-%M-%D',true,'default')" readonly /> <!-- <input name="actualDate" id='actualDate' type="text"
class="searchinput" size="15" readonly /> <a href="javascript:show_calendar('actualDate')"
onMouseclick="MM_openBrWindow('show_calendar','','')" style='text-decoration: none;color: #305C83;' > <img
src="../../images/i_date_normal.gif" width="30" height="18" align="absmiddle" border='0' /> </a> --> </span><span class="search"> <img
src="../../images/chaxun.gif" width="47" height="21" align="absmiddle" onclick='ondate()' style='cursor: hand;' /></span>
</td> </tr> </table></td> </tr> <tr> <td valign="top" height="390"> <table
width="90%" border="0" align="center" cellpadding="0" cellspacing="0"> <tr>
<td> <TABLE cellSpacing=1 cellPadding=0
width="100%" align=center border=0> <TD> <!-- 日期控件 --> <div id="textArea"
style="font-size:14px;"></div> <div id="textArea1" style="font-size:14px;"></div> </TD>
</TABLE> <BR> <table><tr><td><font style='color:red'>*</font>
```

数据来源：中国金融期货交易所网站、广发证券发展研究中心

图4：某动态网页的真实源代码

align="center">持买单量排名</td><td style="color: rgb(35, 102, 164);" class="table_first" colSpan="4" align="center">持卖单量排名
</td></tr><tr vAlign="middle" align="center" bgColor="#f9f9f9"><td style="color: rgb(35, 102, 164);" class="table_first">名次</td><td
style="color: rgb(35, 102, 164);" class="table_first">会员简称</td><td style="color: rgb(35, 102, 164);" class="table_first">成交量</td><td
style="color: rgb(35, 102, 164);" class="table_first">比上交易日增减</td><td style="color: rgb(35, 102, 164);" class="table_first">名次</td><td
style="color: rgb(35, 102, 164);" class="table_first"> 会员简称 </td><td style="color: rgb(35, 102, 164);"
class="table_first">持买单量</td><td style="color: rgb(35, 102, 164);" class="table_first">比上交易日增减</td><td style="color: rgb(35, 102,
164);" class="table_first">名次</td><td style="color: rgb(35, 102, 164);" class="table_first">会员简称</td><td style="color: rgb(35, 102, 164);"
class="table_first">持卖单量</td><td style="color: rgb(35, 102, 164);" class="table_first">比上交易日增减</td></tr><tr vAlign="middle"
bgColor="#ffffff"><td style="word-break: break-all;" class="tb_content_right" align="right">1</td><td style="word-break: break-all;"
class="tb_content_right" align="center">0001-国泰君安</td><td style="word-break: break-all;" class="tb_content_right"
align="right">47499</td><td style="word-break: break-all;" class="tb_content_right" align="right">-9841</td><td style="word-break:
break-all;" class="tb_content_right" align="right">1</td><td style="word-break: break-all;" class="tb_content_right" align="center">0001-国泰
君安</td><td style="word-break: break-all;" class="tb_content_right" align="right">3358</td><td style="word-break: break-all;"
class="tb_content_right" align="right">-14</td><td style="word-break: break-all;" class="tb_content_right" align="right">1</td><td
style="word-break: break-all;" class="tb_content_right" align="center">0018-中证期货</td><td style="word-break: break-all;"
class="tb_content_right" align="right">8686</td><td style="word-break: break-all;" class="tb_content_right" align="right">-56</td></tr><tr
vAlign="middle" bgColor="#f9f9f9"><td style="word-break: break-all;" class="tb_content_right" align="right">2</td><td style="word-break:
break-all;" class="tb_content_right" align="center">0011-华泰长城</td><td style="word-break: break-all;" class="tb_content_right"
align="right">47222</td><td style="word-break: break-all;" class="tb_content_right" align="right">-5658</td><td style="word-break:
break-all;" class="tb_content_right" align="right">2</td><td style="word-break: break-all;" class="tb_content_right" align="center">0011-华泰
长城</td><td style="word-break: break-all;" class="tb_content_right" align="right">2846</td><td style="word-break: break-all;"
class="tb_content_right" align="right">173</td><td style="word-break: break-all;" class="tb_content_right" align="right">2</td><td
style="word-break: break-all;" class="tb_content_right" align="center">0001-国泰君安</td><td style="word-break: break-all;"
class="tb_content_right" align="right">5740</td><td style="word-break: break-all;" class="tb_content_right" align="right">-45</td></tr><tr

数据来源：中国金融期货交易所网站、广发证券发展研究中心

读取了HTML网页源代码后，我们需要提取其中的正文信息和其它有用信息，我们用的方法大致如下：

1. 缩小范围，找到有用信息所在区域；
2. 找到标题、网址和日期等信息所在位置的特征，然后记录这些信息；
3. 找到正文以及相应段落所在位置的特征，然后提取正文。

我们以某网页为例来展示文本提取的过程，图5是该网页的源代码，可以看出，其中的有用信息可以轻易地被提取出来。

图5：文本提取的过程

数据来源：全景网、广发证券发展研究中心

最后我们来看PDF文本的提取方法，我们编写这样一个MATLAB函数，输入参数为PDF文件的网址，输出结果为纯文本格式的字符串，其主要步骤如下：

1. 使用wget下载PDF文件；
2. 使用工具将PDF文件转换为TXT文件；
3. 对TXT文件进行细节上的处理，并判断转换是否成功；
4. 打开TXT文件，读入MATLAB，并输出。

我们以某PDF文件为例来展示PDF文本提取的过程，图6为该PDF文件的原文，图7为PDF文本提取后的内容。

图6：某PDF文件的原文

股票简称：浦发银行　　股票代码：600000　　编号：临2012-009

上海浦东发展银行股份有限公司

关于花旗银行海外投资公司转让本公司股份的公告

本公司及董事会全体成员保证公告内容不存在虚假记载、误导性陈述或者重大遗漏，并对内容的真实性、准确性和完整性承担个别及连带责任。

上海浦东发展银行股份有限公司（以下简称“公司”）于2012年3月19日接到公司股东花旗银行海外投资公司（以下简称“花旗银行”）通知：花旗银行于2012年3月19日通过大宗交易的方式向几家机构投资者转让了506,164,207股的本公司股份，占本公司已发行总股份的2.714%。至此，花旗银行在本公司的股份为零。

特此公告。

上海浦东发展银行股份有限公司

二〇一二年三月十九日

数据来源：上海证券交易所网站、广发证券发展研究中心

图7：PDF文本提取后的内容

股票简称：浦发银行　股票代码：600000　编号：临2012-009↓
上海浦东发展银行股份有限公司↓
关于花旗银行海外投资公司转让本公司股份的公告↓
关于花旗银行海外投资公司转让本公司股份的公告↓
本公司及董事会全体成员保证公告内容不存在虚假记载、误导性陈述或者重↓
大遗漏，并对内容的真实性、准确性和完整性承担个别及连带责任。↓
上海浦东发展银行股份有限公司（以下简称“公司”）于2012↓
年3月19日接到公司股东花旗银行海外投资公司（以下简称“花旗银↓
行”）通知:花旗银行于2012年3月19日通过大宗交易的方式向几家机↓
构投资者转让了506,164,207股的本公司股份，占本公司已发行总股↓
份的2.714%。至此，花旗银行在本公司的股份为零。↓
特此公告。↓
上海浦东发展银行股份有限公司↓
二〇一二年三月十九日↓
1←

数据来源：上海证券交易所网站、广发证券发展研究中心

四、文本脱水和模式识别

将PDF转换为纯文本的时候，无法提取PDF文件中的格式和排版，因此，提取后的文本不可避免的存在一部分无用信息。文本脱水是指去除文本中的无用信息，文本脱水步骤中的细节很多，其主要想法是，首先尽可能不删除有用信息，然后尽可能删除无用信息。例如，对于卖方研究报告来说，无用信息包括图表中的数字和分析师信息等。

我们以某研究报告为例来展示文本脱水的过程，图8为该研究报告的纯文本原文，图9为该研究报告文本脱水后的内容。

图8：某研究报告的纯文本原文

制造业依然处于较低迷的状况↓
（1）从汇丰制造业 PMI 来看，12 月 PMI 总指数为 48.7（11 月为 47.7），很明显处于继续萎缩的状况。汇丰认↓
为主要原因是：需求低迷，新订单连续两月下滑；海外需求疲软，新出口订单指数持续萎缩。↓
（2）从中国制造业 PMI 来看，12 月 PMI 总指数为 50.3（11 月为 49），处于微弱的扩张状态。主要原因是生产↓
指数从 11 月的 50.9 上升至 12 月的 53.4。但实际上，我们经过季节调整发现，12 月 PMI 总指数为 49.97（11 月↓
为 58.67），生产指数为 52.05（11 月为 50.84）。由此可以认为，12 月制造业总体状况依然较低迷。另外，12↓
月新订单与新出口订单指数分别为 49.8 和 48.6，处于明显的萎缩状态。外需减弱导致新订单减少，将会使生产下↓
降，积压存货增加，使整个制造业陷入较低迷的状态。这与汇丰制造业 PMI 数据显示的结果是一致的。↓
非制造业 PMI 明显季节性回升↓
中国非制造业 PMI 中，12 月商务活动指数为 56.0（11 月为 49.7）。我们认为，这里面包含有明显的季节性因素，↓
从图 3 可以明显看出这种季节性。而且，图 4 也显示出外需乏力对非制造业 PMI 订单也产生了明显影响。↓
图 1：制造业 PMI 订单指数继续萎缩
图 2：汇丰制造业 PMI 与国家统计局制造业 PMI 走势↓

数据来源：迈博汇金网站、广发证券发展研究中心

图9：文本脱水后的内容

外需乏力的影响正在深入020-87555888-8666执业编号：S0260511020003制造业依然处于较低迷的状况（1）从汇丰制造业PMI来看，12月PMI总指数为48.7（11月为47.7），很明显处于继续萎缩的状况。汇丰认为主要原因是：需求低迷，新订单连续两月下滑；海外需求疲软，新出口订单指数持续萎缩。↓
（2）从中国制造业PMI来看，12月PMI总指数为50.3（11月为49），处于微弱的扩张状态。主要原因是生产指数从11月的50.9上升至12月的53.4。但实际上，我们经过季节调整发现，12月PMI总指数为49.97（11月为58.67），生产指数为52.05（11月为50.84）。由此可以认为，12月制造业总体状况依然较低迷。另外，12月新订单与新出口订单指数分别为49.8和48.6，处于明显的萎缩状态。外需减弱导致新订单减少，将会使生产下降，积压存货增加，使整个制造业陷入较低迷的状态。这与汇丰制造业PMI数据显示的结果是一致的。↓
非制造业PMI明显季节性回升中国非制造业PMI中，12月商务活动指数为56.0（11月为49.7）。我们认为，这里面包含有明显的季节性因素，从图3可以明显看出这种季节性。而且，图4也显示出外需乏力对非制造业PMI订单也产生了明显影响。↓
图1：制造业PMI订单指数继续萎缩图3：非制造业PMI商务活动指数存在季节性广发证券公司或其关联机构可能会持有报告中所涉及的公司所发行的证券并进

数据来源：迈博汇金网站、广发证券发展研究中心

模式识别和文本脱水所用的方法类似，都需要对文本进行模式匹配，区别在于，文本脱水是删除符合特定模式的文本，而模式识别是找出符合特定模式的文本。

例如图10这段话，我们可以看出这是对CPI的预测，而预测的观点出现在最后一句："预计12月CPI持平于上月，即为4.2%"。这里面，"CPI"字样告诉我们这是对CPI的预测，"预计"这个词明确了这是一次预测而不是总结，"12月"字样告诉我们这是对12月CPI的预测，而报告日期是2012年1月6日，因此这是对2011年12月CPI的预测。"持平于上月"在这里的作用不大，"即为4.2%"是观点，因此，我们从这段话中知道了，某分析师对2011年12月的CPI预测值为4.2%。

图10：模式识别举例1

物价方面，11月CPI出现较为明显的回落，但12月以来农产品价格由于季节因素再度反弹，从商务部以及农业部监测的农产品价格来看，蔬菜类涨幅较为明显，而肉类，特别是猪肉价格亦止跌回涨，物价上涨在春节前或仍将维持较大压力。预计12月CPI持平于上月，即为4.2%。↓

数据来源：迈博汇金网站、广发证券发展研究中心

又比如图11这段话，我们可以看出其中包含了对工业增加值的预测，预测的观点为："我们预计12月份工业增加值同比增幅将从11月份的12.4%小幅降至12.3%"。这里面，出现了12月和11月两个月份，也出现了两个数值。在进行模式识别时，我们需要识别出，某分析师对2011年12月工业增加值的预测为12.3%。

图11：模式识别举例2

作者对北京大学教授、北京高华证券有限责任公司特别顾问12月份数据前瞻：GDP环比增幅和货币/信贷增速反弹，通胀率保持低位各政府部门将从1月9日开始公布2011年12月份经济数据和四季度GDP数据，具体时间见图表1。我们对这些数据的预测及观点如下：2011年四季度GDP环比增速反弹我们预计2011年四季度GDP同比增速从2011年三季度的9.1%放缓至8.8%，对应季调后折年季环比增幅为8.8%，高于2011年三季度的8.5%。2011年全年GDP增速有望达到9.2%，略高于我们9.1%的预测。GDP增幅强于预期受益于自10月底开始的政策放松和好于预期的出口增长。此外，今年粮食产量大增4.5%有望推动第一产业增长。虽然农业生产贯穿全年，但粮食丰收的主要影响体现在四季度GDP数据我们预计12月份工业增加值同比增幅将从11月份的12.4%小幅降至12.3%。隐含的季调后月环比折年增幅将从11月份的13.0%升至16.1%。↓

数据来源：迈博汇金网站、广发证券发展研究中心

广发金融工程研究小组

罗军，首席分析师，华南理工大学理学硕士，2010 年进入广发证券发展研究中心。
俞文冰，首席分析师，CFA，上海财经大学统计学硕士，2012 年进入广发证券发展研究中心。
叶涛，资深分析师，CFA，上海交通大学管理科学与工程硕士，2012 年进入广发证券发展研究中心。
安宁宁，资深分析师，暨南大学数量经济学硕士，2011 年进入广发证券发展研究中心。
胡海涛，分析师，华南理工大学理学硕士，2010 年进入广发证券发展研究中心。
夏潇阳，分析师，上海交通大学金融工程硕士，2012 年进入广发证券发展研究中心。
蓝昭钦，研究助理，中山大学理学硕士，2010 年进入广发证券发展研究中心。
李明，研究助理，伦敦城市大学卡斯商学院计量金融硕士，2010 年进入广发证券发展研究中心。
史庆盛，研究助理，华南理工大学金融工程硕士，2011 年进入广发证券发展研究中心。
汪鑫，研究助理，中国科学技术大学金融工程硕士，2012 年进入广发证券发展研究中心。
谢琳，研究助理，上海交通大学金融学博士研究生，2011 年进入广发证券发展研究中心。

相关研究报告

	广州市	深圳市	北京市	上海市
地址	广州市天河北路 183 号大都会广场 5 楼	深圳市福田区民田路 178 号华融大厦 9 楼	北京市西城区月坛北街 2 号月坛大厦 18 层	上海市浦东南路 528 号上海证券大厦北塔 17 楼
邮政编码	510075	518026	100045	200120
客服邮箱	gfyf@gf.com.cn			
服务热线	020-87555888-8612			

免责声明