“智慧政务”中的文本挖掘应用

摘要：近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步普及他们成为政府了解民意的重要渠道，各类社情民意相关的留言数据量不断攀升，所以像以往一样主要靠人工来进行留言划分和热点整理的相关部门的工作已经不再适用。而且，随着大数据、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。针对以上背景我们需要解决如下问题。

通过分析可以看出问题是层层递进的，所以我们通过五个步骤对问题逐一进行解决。问题一是工作人员在处理网络平台的群众留言时，首先会按照一定的划分体系对留言进行分类,以便后续将群众留言分派到相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。我们需要根据已给出的数据，建立关于留言内容的一级标签分类模型。我们通过对数据的预处理，结合 VIPS 算法并进一步挖掘分类得到分类标签。

问题二是对“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”这一具体热点问题进行分析。我们需要根据提供的数据将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。我们通过数据筛选得出群众留言的集中问题，并绘制出热点问题前五名及其相关留言问题的表格,结合表格进行分析。并结合情感分析。了解留言中的情感因素，寻找他们想要了解的内容，并通过评分的方式让用户了解政府对城市发展规划；本文基于中文文本情感分析，获取留言者者情感，分析文本中的情感倾向，并构建特征词表，进而绘制“留言-留言特征”二分网络，分析每条评论中的情感特征，结合用户的评价评分信息，最终构建包含留言情感信息的规划建设模型。通过与城市建设初期的规划比较，发现在考虑到市民情感信息后的意见与最初规划有一定差异，差异大部分是由于政府规划建设未能满足市民要求导致的，据此给出评价模型。

我们的工作主要在于对市民提出的问题能跟高效的分配到各部门中及时进行回复。针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。我们首先分析了相关部门对问题的回复，由所给数据我们发现目前的处理方法依旧未能实现高效，还有很多种分配不属于部门的情况出现，针对这一问题我们要更加完善模型，而且在回复市民问题时不能简而概之，要给出合理具体的方案并加快实施进度。

关键词：VIPS 算法，达闻微指数，二分网络分析

目录

1、挖掘目标

近年来，互联网络作为一个正在加速膨胀的思想阵地，已越来越引起足够的重视与运用，互联网的信息繁杂多样，良莠不齐，进步健康有益的信息大量涌现。同时，反动、迷信、黄色的信息也随之而来。加之互联网信息的虚拟性、隐蔽性、发散性、渗透性和随意性等特点，越来越多的人们愿意通过这类渠道表达自己的个人想法，于是互联网成为了一个社情民意的重要来源，对于互联网上社情民意的调查、分析与监测非常重要。本次建模为了提升政府对各类社情民意的管理水平和施政效率，采用 VIPS 算法深度学习模型的数据挖掘方法，达到以下目标：

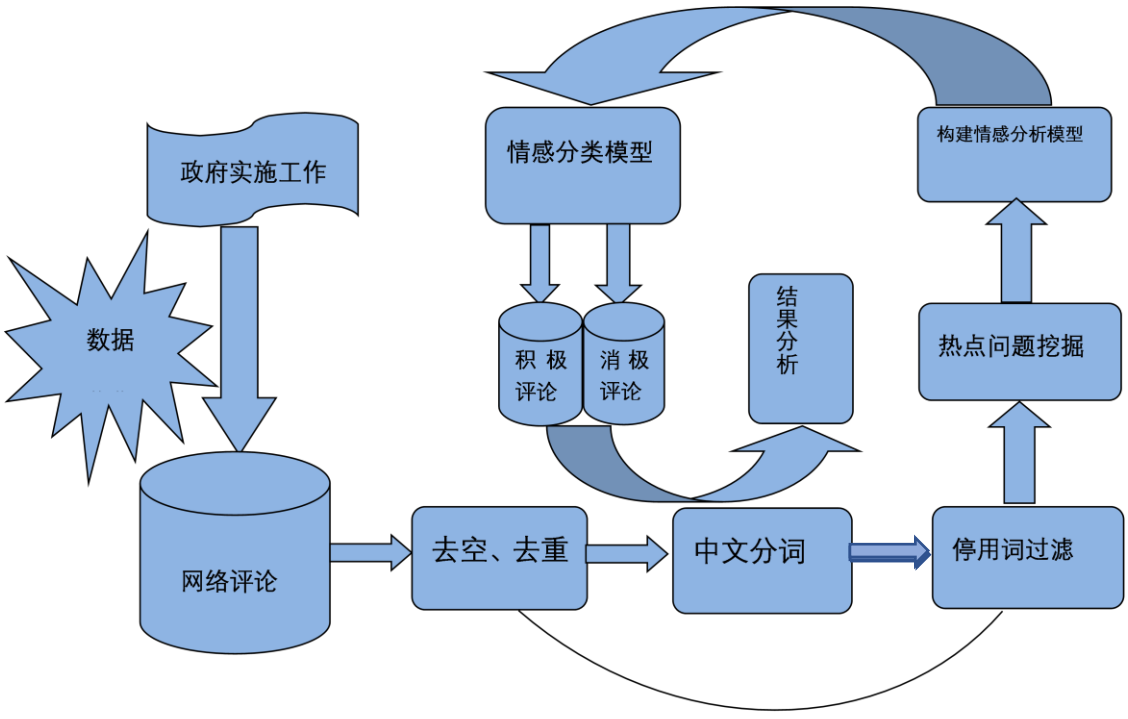
（1）工作人员在处理网络平台的群众留言时，首先会按照一定的划分体系对留言进行分类，以便后续将群众留言分派到相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。我们需要根据已给出的数据，建立关于留言内容的一级标签分类模型。

（2）对“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”这一具体热点问题进行分析。我们需要根据提供的数据将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。

（3）要通过相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

2、分析方法与过程

2.1 总体流程



本用例主要包括以下几个步骤：
步骤一：分析提供的评论数据，对评论数据分析的是本次数据挖掘分析的第一步。由于本题已将所需数据给明，所以我们需要建立可行的模型先对数据进行大致分类在进行分析。要明确问题及结合可行的数学模型思考分类标准的可行性。
步骤二：数据预处理。第一步要“去空、去重”；第二步对评论数据进行中文分词，将一句评论分成多个词语进一步分析：第三步进行停用词过滤，去除掉评论中与情感判定不相关的词。找出可作为分类的标准。
步骤三：热点问题挖掘。及时发现问题并对热点问题进行回应，定义合理的热度评价指标，并给出评价结果。
步骤四：情感分析。针对事先指定的分析对象，系统自动分析海量文档的情感倾向：情感极性及情感值测量，并在原文中给出正负面的得分和句子样例。利用构建的模型分析得出评论数据的情感倾向。
步骤五：属性提取并统计。将所有提及到某一分类标准的评论数据从实验数据集中筛选出来，针对相关部门对这一分类留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

2.2 具体步骤

2.2.1 步骤一：题目已将社情民意分逐渐细为三级标签体系。我们可以看出一级标签就可以让问题分到不同的管理结构。如下我们借助 VIPS 算法说明其可行性。
① Web 的基于视觉的内容结构描述
我们提出 VIPS(Vision-based page segmentation)算法用以提取给定网页的语义结构。这种语义结构是层次性的结构，在该结构中，每一个结点代表一个语义块。每一个语义块都定义一个 DOC 值来描述该语义块内部内容的关联性。DOC 的值越大，则表明语义块内部的内容，它们之间的联系越紧，反之越松散。VIPS 算法充分利用了 Web 页面的布局特征：它首先从 DOM 树中提取出所有的合适的页面块，然后根据这些页面块检测出它们之间的所有的分割条，包括水平和垂直方向。最后基于这些分割条，Web 页面的语义结构将被重新构建。对于每一个语义块又可以使用 VIPS 算法继续分割为更小的语义块。因此整个 VIPS 算法时自顶向下，非常高效的。

VIPS 算法中首先也定义了“基本对象”的概念，通常 DOM 树上的叶子结点被定义为基本对象，因为这些结点已经不能再被继续分割了。在本论文中，我们首先引入了基于视觉的内容结构，它里面的每一个结点我们称之为“块”，这些块或者是一个基本对象或者是一些基本对象的组合。有一点需要注意的是，基于视觉的内容结构中的块与 DOM 树中的结点没有绝对的对应关系。
VIPS 算法中 Web 页面的结构定义如下。

对于每一个页面而言，我们可以将其看作一个三元组$\Omega=(O,\ \phi,\ \delta)$，其中$O=(\Omega_1,\Omega_2\ldots,\Omega_N)$，表示给定页面上的所有的语义块的集合，这些语义块之间没有重叠覆盖，而每一个语义块Ω_i又可以被定义为前面所描述的三元组$\Omega_i=(O_i,\ \phi_i,\ \delta_i)$，如此迭代循环；

$\phi=(\delta_1,\delta_2,\ldots,\delta_t)$，表示当前页面上的所有的分隔条的集合。事实上，一旦确定了一个页面上的两个语义块，那么这两个语义块之间的分隔条也就被确定了。当然，VIPS 中的分隔条并不是真正存在的分隔条，而是虚拟。分隔条包括水平分隔条，也包括垂直分隔条。每一个分隔条都具有一定的宽度和高度。

$\delta=(\zeta_1,\zeta_2,\ldots,\zeta_M)$则描述了$\Omega$集合中两个语义块之间的关系，这种关系可以用下面的式子描述：$\delta=O\times O\to\phi\cup\{NULL\}$其中的每个$\zeta$都是一个形如$(\Omega_i,\Omega_j)$二元组，其表示块$\Omega_i$和$\Omega_j$之间存在一个分割条。

比如，对于 VB2，从它的内部又可以检测出三个子对象和两个分隔条。

对于每一个 Block，VIPS 算法都定义一个 DoC(Degree of Coherence)与之对应。该值的大小反映了当前语义块内部内容联系的紧密程度。如果，它具有下面两个重要的特性：

1)DoC 的值越大，则语义块内部的内容之间的联系紧密程度就越大，它们之间就关系就越连续，反之越小。

2)在层次数上，语义块的子块的 DoC 的值肯定要比父块的值大。

在 VIPS 算法中，DoC 的值位于 1 到 10 之间。不过这个范围是可以更改的。在对 Web 页面进行语义分割之前，我们首先设定一个预定义 DoC 值 PDoC(Permitted Degree of Coherence)，通过该值来限定分割的语义块的粗糙程度。当语义块的 DoC 值达到 PDoC 之后，迭代分割就停止。PDoC 越小，则分割的语义块就越粗糙，反之，分割的语义块就越精细。不同的应用程序可以设置不同的 PDoC 值来达到自己的要求。

基于视觉的页面分割最主要的目的就是对给定的页面进行语义分割，因此分割后生成的基于视觉的内容结构中的结点通常总是一定的语义单位,包含一定的语义。

② VIPS 算法描述

这部分我们将详细介绍 VIPS 算法。整体来说，页面的基于视觉的内容结构是结合 DOM 树以及一些视觉提示信息而得到的。它具有三个步骤；页面块提取、分隔条提取以及语义块重构。这三个步骤联合一起作为一次语义块检测的完整步骤。Web 页面首先被分割为几次比较大的语义块，同时这几个语义块所组成的层次结构将被记录下来。对于检测出来的每一个大的语义块分页过程又可以继续进行，直到语义块的 DoC 值达到预先设定的 PdoC 为止。

在每次迭代循环中，当前逻辑块的 DOM 树结构以及它的视觉信息都将被获取。然后，从 DOM 树的根结点开始，逻辑块检测过程将基于视觉信息开始从 DOM 树中开始检测页面块。每一个 DOM 结点都会被检查它能够构成一个单独的页面块。如果不能，那么它的子结点将被执行同样的检查。对于每一个提取出来的页面块，我们都会根据当前页面块的内部可视属性赋予一个 DoC 值。当本次迭代过程中所有的页面块都被检测出来之后，它们将被保存到页面块池中。基于这些页面块，分隔条检测过程将开始工作。这些页面块之间的所有的水平分隔条和垂直分隔条最终将被识别出来并且赋予一定的宽度和高度。基于这些分隔条，页面的布局层次将被重新构建——一些页面块将被合并，形成语义块。最终，本次迭代过程中的所有语义块都被检测出来。

迭代过程是否需要继续进行取决于本层次的语义块中是否存 DoC 值小于 PdoC 的语义块。对于那些$DoC\geq PdoC$的语义块，分隔过程将停止，否则分隔过程将继续。当所有的语义块被提取出来后，最终整个 Web 页面的基于视觉的内容结构也就构建完成。

③ 语义块提取及分隔条检测

（1）语义块提取

在这步骤中，我们的目标是提取出当前子页面中所包含的所有的可视语义块。通常情况下，DOM 树中的每一个结点都可以表示一个可视语义块。不过，在 HTML

中，一些标签比如<TABLE>和<P>通常用来进行数据组织，因此不适合表示单独的可视语义块。对于这种结点，对他们的提取将被它们的孩子结点替代。而且由于HTML 语法的灵活性，很多的 Web 页面并没有严格遵循 W3C 的 HTML 规范，这导致DOM 树并不能总是能反映不同的 DOM 结点之间的关系。对于提取出来的每一个可视语义块，我们将根据它的内部的视觉差异设置它的 DoC 值。整个迭代提取过程可以用下面的算法描述：

```
Algorithm DivideDomtree(pNode,nLevel)
{
  IF (Dividable(pNode,nLevel)==TRUE)}
    FOR EACH child OF pNode{
       DivideDomtree(child,nLevel);
    }
   }ELSE{
    Put the sub-tree(pNode)into the
        pool as a block;
   }
}
```

如何判断给定的结点能否被继续分割，我们给出下面的几个方面进行判断:
DOM 结点本身的属性。比如当前 DOM 结点的标签，结点的背景色，当前结点所代表的页面块的大小，形状。
当前 DOM 结点的孩子结点。比如孩子结点的标签，孩子结点所代表的区域的背景色，前景色，区域的大小以及不同类型的孩子的数目等等。

基于 WWW HTML 规范 4.0，我们将 DOM 结点分为两大类：inline 结点和 line-break 结点。
所谓 Inline 结点是指:如果该结点的标签能够影响文字的外观同时不会引起换行的话，那么这类结点我们称之为 Inline 结点，比如<B>、<BIG>、<EM>、<FONT>、<STRONG>、<U>等等，这类结点通常仅仅影响文字的外观而不会影响文字的布局。所谓 Line-break 结点，则就是除了 inline 结点之外的所有结点。
（2）分隔条检测

当所有的页面块被提取出来之后，它们都被保存在页面块池中以便进行分隔条检测。在 VIPS 算法中，分隔条是 Web 页面中的垂直的或者水平的行。从视觉的角度而言，separators are good indicators for discriminating different semantics within the page。

在 VIPS 中，一个可视的分隔条可以用二维向量(Ps，Pe）描述，其中，Ps 是分隔条的起始坐标,而 Pe 则是分隔条的终止坐标。坐标的单位全部为像素 pixel。根据 Ps 和 Pe，很容易计算当前分隔条的宽度和高度。
分隔条的检测算法如下描述：
1) 初始化分隔条列表。最早的分隔条列表中仅仅存在一个分隔条，它的起始止坐标为（Pbe,Pee），分别对应整个 Web 页面的起始坐标和终止坐标。
2）对于页面块池中的每一个页面块，它与分隔条的关系包括下面三种：
页面块被包含在分隔条中,此时,该分隔条将从页面块的边缘裂变为多个分隔条；页面块与分隔条发生部分重合，那么根据页面块的边界重新调整分隔参数；页面块跨越分隔条，那么此时移除该分隔条。
3) 移除页面边缘的四个分隔条。

分隔条通常用于区别不同语义的页面块，因此基于给定分割条两边的语义块的在视觉上的差异，我们还可以设置分割条的权重。如果分隔条的权重越重，该分隔条最终成为分隔条的可能性就越大。

2.2.2 步骤二：与数据库中的结构化数据相比，从网页上选取的数据属于半结构化或者非结构化数据，即具有有限的结构，或者根本就没有结构，即使具有一些结构，也是着重于格式，而非文档内容，不同类型文档的结构也不一致。此外，网页数据缺乏机器可理解的语义，而数据挖掘的对象局限于数据库中的结构化数据，并利用关系表格等存储结构来发现有价值的信息,因此有些数据挖掘技术并不适用于网络文本挖掘，即使可用也需要建立在对网络文本数据进行预处理的基础之上。如果要对网络评论数据进行情感分析，就必须先将文本数据进行预处理，转化为结构化的数据。该步骤中，从以下几个方面对步骤一中从网页上爬取的评论数据进行预处理。

1)“去重”、“去空”

对于存储了全部网络商业评论的 txt 文件，每行代表了一个评论文本但是难免会出现两个完全一样的文本和一些空行。所以本文首先进行了“去重”、“去空”的预处理工作。

在导入评论文本时，同时进行了是否为空的判断，只导入不为空的文本，从而过滤掉了空白文本，“去空”的程序段如图下所示：

```
StreamReader sr=new StreamReader;
String line;
while((line =sr.ReadLine( ))!=null)
 {
   if(line.ToString( )!="") （%去掉空文本）
    {
       CommentsList.Add(line.ToString( ));
    }
}
```

将非空的评论文本导进 List 后，再进行去除重复处理，过滤掉重复的评论文本，“去重”的程序段如下：

```
CommentsList2.Add(CommentsList[0]);
for(int i=1;i<CommentsList.Count;i++)
 {
     IsRepeated=false;
     for(int j=0;j<i;j++)
     {
         if(CommentsList[i].Equals(CommentsList[j]))
         {
             IsRepeated=true;
             break;
         }
     }
     if(!IsRepeated)
     {
         CommentsList2.Add(CommentsList]);
```

```
        }
    }
```

2)中文分词

中文分词(Chinese Word Segmentation)，也可称为中文切词，指的是通过某种特定的规则，将中文文本切分成一个一个单独的词。本文使用 NLPIR 汉语分词系统（又名 ICTCLAS2015）进行分词，它是中科院张华平博士主持开发的中文汉语分词工具，主要功能包括中文分词；词性标注；命名实体识别；用户词典功能；支持 GBK 编码、UTF8 编码、BIG5 编码。新增微博分词、新词发现与关键词提取功能。本文用到了在 NLPIR 官网上下载到的 NLPIR.dll 程序包，在 Microsoft Visual Studio 2012 编程环境中用 C#高级语言程序对 NLPIR.dll C++程序包进行调用，实现对网络商业评论文本进行批量分词处理和词性标注。主要程序段如下所示：

```
    if(! NLPIR_Init("F:/ICTCLAS2015",0, " "))
    {
            System.Console.WriteLine("InitICTCLASfailed! ");
            return;
}
else
            System.Console.WriteLine("InitICTCLAS success!");
Console WriteLine();
System Console WriteLine("分词处理中…");
for (int i=0;i<content Count;i++) /
{
    IntPtr intPtr=NIPIR_Paragraph Proccss(content[i]);
    String str=Marshal.PtrToStringAnsi(inPtr);
    Content_seg.Add(str)
}
```

3)停用词过滤

评论文本在经过去重、去空、中文分词后，并非所有的剩下的词语都可以作为特征词，里面还有一些包含的信息量很低甚至没有信息量的词语，需要将它们过滤掉，否则将会影响下文的分析的正确率。在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言之前会自动过滤掉某些字或词，这些字或词即被称作 Stop Words（停用词）。

本文采用了“词性+停用词表”的过滤方法。在上文已经提到了中文分词后的词语还带有词性的标注，所以本文根据中科院《计算所汉语词性标记集》将上述停用词词性都写进 StopwordPropsList 里面，部分如下所示，然后对每个分词后的文本进行遍历扫描，把对应词性的词语全部过滤掉。

```
//介词
StopwordPropsList.Add("p");
StopwordPropsList.Add("pba"); //把
StopwordPropsList.Add("pbei"); //被
//连词
StopwordPropsList.Add("c");
StopwordPropsList.Add("cc"); //并列连词
```

为了把评论文本中包含的停用词过滤干净，本文还利用了《哈工大停用词表》进行辅助过滤，在词性过滤后再把文本中存在于停用词表的词语过滤掉，进一步过滤掉评论文本中的停用词。

2.2.3 步骤三：热点问题挖掘

①对学科热点及其变化轨迹进行分析，可以了解该学科过去和现在的热点及未来发展趋势，因此热点研究一直受到人们的关注。传统研究方法主要通过电子期刊、硕博论文、立项基金或引文信息源，采用词频分析、共词分析、聚类分析等方法，利用已有的或自行开发的软件进行分析挖掘。传统信息源具有高质量和权威性的特点，在此基础上进行热点研究也具有相当的可信度和说服力，但传统信息源最大的缺点之一就是更新滞后，目前一篇文章从完成到正式发表大约历时一年左右。信息时代技术瞬息万变，根据传统数据源得来的学科热点有时只能代表过去，大大降低了学科热点研究的质量，偏离了研究的原始目标。微博等网络手段是一个基于用户关系的信息分享、传播及获取平台，用户可以通过及各种客户端组建个人社区，以一定数目的文字或者图片、视频、网页链接等方式更新信息，并实现即时分享。下面我们以从微博挖掘数据为例详细分析数据挖掘的方法。微博微博的最大特点就是快速更新、群策群力，是否可以根据微博的特点来获得学科前沿资料并对学科热点进行发现、追踪与分析，弥补传统学科热点分析的不足之处，将是下文的研究目标。

随着 Web2.0 的出现，互联网用户行为已发生了很大变化，从微博上获得信息较以往变得容易许多，基于微博的各种应用不断涌现，并逐渐为企业、政府、出版社等部门所认可。目前微博应用领域主要有：企业应用、新闻媒体、政府舆情、学术出版等，但将微博用于学术领域的目前尚未可见，其他领域应用中所涉及的技术方法和思路对微博在学术领域的应用研究将有很大的参考借鉴作用。Xi下文基于通过微博平台了解他人研究进展，获取学科前沿信息，预测学科热点具有可行性，理由如下：微博上包含了大量专家学者、研究人员；微博学术信息更新快，信息可公开获取；微博的分众聚类可形成学术圈；第三方应用让学术信息容易被计算机自动获取和分析。

②基于微博预测学科热点的步骤

热点发现：虽然微博中存在大量学术前沿信息，但这些信息分散在微博的各个角落，如何把他们汇聚起来从中发现学科热点是该步骤要完成的任务。

信息采集和预处理。信息的采集是热点发现的第一步，微博信息的采集主要分两个方面：微博内容；作者 ID。

微博内容的采集可用作热点挖掘的语料，而微博作者 ID 的采集也同样重要，由于微博存在关注和转发功能，通过分析作者 ID，可以找出其中的学术权威，对于了解学术最新进展有重要意义。采集后的数据还存在大量信息冗余，因此要进行信息的预处理，主要包括合并数据、文本清洗、添加自定义词等步骤。

发现学科热门词汇。将采集到的数据进行特征提取，分类聚类以及共词分析等方法处理可以找出某一领域相关的热门词汇。

发现学术意见领袖。意见领袖是最为活跃的话语群体，他们的意见和观点对事件的发展方向产生了深远影响。在微博的学术领域，意见领袖可以分成两种：一种是以知名学者为代表的学术权威，另一种则是从学术研究或工作实践出发，不断提供最新学科前沿信息的知识工作者。发现学术意见领袖可采用社会网络分析、聚类等方法。

热点追踪。在热点发现中获得的学术热点和学术领袖只能代表过去，如何沿着这

些热点继续追踪，不断了解学术发展的现状，并对未来发展做出预测是该步骤真正要完成的工作。

热点关键词微博跟踪。关键词跟踪有两种方法：①使用微博平台自带的检索工具；②使用第三方软件。自带检索工具使用简便，缺点是要靠使用者人工刷新才能获得最新数据，无疑会耗费过多的人力成本。目前已有第三方软件弥补这一缺陷，如针对新浪微博的“热点搜索工具”安装后可订阅某关键词实时更新，大大方便了使用者对学术热点的追踪。

关注学术领袖。微博特有的关注功能可以跟踪订阅他人的最新发言，了解他人的学术观点和学术进展，从而形成虚拟学术圈。除此之外，还可以采用第三方软件，如绘制学术圈关系图等。

其他辅助工具。微博学术信息的特点是更新快，有利于了解学术最新动向，但由于字数限制，很多问题无法进行详细完整的描述，因此微博学术研究还要借助其他工具。例如通过搜索引擎和期刊数据库等工具获得更为详尽的学术资料，这是微博本身无法实现的。

热点分析。获得学术热点以及相关资料后，除了可以采用传统分析软件进行分析外，目前针对微博开发的第三方分析应用软件也大量涌现，这类软件的特点是实用性强，样式丰富，使用简便，可以方便地对微博数据进行分析。由于其设计时考虑了微博信息的特点，因此具有传统分析软件无法比拟的优势，例如对转发深度的分析、对粉丝标签的分析、对社交网络的绘制、对作者地域的统计，等等。灵活地应用这些分析软件，我们可以对热点进行比较研究，找出热点作者的学术交际圈，了解学术意见领袖关注的话题，对学科发展趋势和发展轨迹进行预测。

现如今科技发展迅速，关于热点数据挖掘的方式方法还有很多种，我们也可以通过多种渠道获取数据，这就更使得我们建立的模型适应性更强。关于我们所列举的微博数据挖掘方法既有他的适用性也有其局限性。他作为目前比较大的信息交流平台用户很多，信息传播迅速，但其使用年龄并未能包含全部这可能就会导致数据的统计出现纰漏。但随着科技普及应该会有越来越多的用户加入其中，并且我们可以通过从多方面收取建议来完善政府的服务工作。

③基于新浪微博的学科热点研究实例——以数据挖掘领域为例

目前国内主流微博主要有新浪、腾讯、人民网三大微博运营商，尤其是新浪微博靠其较完善的功能和服务以及名人认证、内部邀请等营销模式，截至 2010 年 10 月底已经有超过 5000 千万用户，成为中国最主流、最火爆的微博产品。本文就以新浪微博为例给出数据挖掘领域热点研究实例。

数据挖掘领域热点发现

信息采集与清洗。首先选取关键词“数据挖掘”使用 ROST CM 软件在规定时间段内采集数据，共采集语料文件约 9 万字左右。由于采集到的原始文件中包含大量冗余数据、异构数据，经过数据合并、文本清洗和添加自定义分词词组后，得到符合要求的语料文件，为下一步的数据分析奠定基础。

特征提取与词频分析。利用特征词词频分析找出围绕“数据挖掘”的相关关键词。这些特征词之间的语义关系怎样，又表达了怎样的含义，接下来将使用共词矩阵和内容语义网络进一步分析。

共词分析。特征词词频仅仅帮助笔者得到了围绕数据挖掘相关的高频词语，但词语本身表达不了完整的含义。

挖掘结果对比与评价。通过以上三个步骤，获得了数据挖掘领域的热点话题，接下来通过同传统研究数据比较后对挖掘结果进行评价。

作者社会网络分析。除了研究关键词热点外，对博文作者的研究也是必不可少的，通过对作者及其学术圈的分析和跟踪可以更深入地了解某个热点话题。

④数据挖掘领域热点分析

除了发现和追踪学术热点外，还可以对结果进一步挖掘，如使用 SPSS 等专业分析软件或是针对微博开发的第三方软件，由于基于专业软件的数据挖掘研究较多，下面将针对两个针对微博的第三方软件案例加以说明。

使用“达闻微指数”应用对热点关键词进行对比分析。达闻微指数是以微博搜索为基础的免费海量数据分析服务，它可反映同一关键词在过去某段时间内每天的变化趋势，还可以反映不同关键词在过去一段时间里的用户关注度。

使用“微博小分析”应用对学术意见领袖进行分析。微博小分析应用可以分析某个作者转发最多的博友、转发最多博友的标签、粉丝博友的标签和关注博友的标签。通过对学术意见领袖转发情况的分析，对最关心的人和最关心的话题标签进行分析，可以按图进一步挖掘出该领域内其他重要学术成员和研究方向。

⑤存在的问题及展望

目前利用微博预测学科热点还处于探索阶段，因此仍然存在许多难点和障碍，主要有：分词词表对结果产生影响较大；评价微博学术权威的标准有待商榷；学术权威的微博开通率影响研究质量；隐性知识不愿意公开等。

2.2.4 步骤四：情感分析

①在信息时代下，越来越多的人通过网络留言等手段表达对政府工作的认可以及存在的问题。特别是随着网络技术的不断发展，出现博客、微博、论坛等众多的网络平台为网络用户提供了更宽阔的平台来交流信息、表达意见。往往这些在线评论的文本信息不仅蕴含着用户的情感态度，也蕴含着巨大的参考价值，其反应社会集体的情感状态，与此同时情感在人类决策时扮演着重要地位。因此，在线评论不仅是了解群众的需求、喜好的重要信息来源，而且也为政府提供了有效了解群众意见的手段和反应城市建设的“晴雨表”。

本文将用户评论信息与评分评价信息相结合，使用文本情感分析的方法，以市民对政府工作的留言信息为语料源，通过用户与评论特征构建的二分网络，分析用户参与评论中的评论倾向，将评论用户与评论特征进行点线连接的方式构建网络联系，进一步提炼用户评论的评论特诊信息以及市民的问题：结合情感分析得到的单句情感得分，得到用户的特征情感得分；利用进一步结合评分评价得分建立新的评分模型，将得到更新的评分与原计划工作效果作比较，提升了政府了解工作进展的层次和效果以及工作结果的有效性和可信性。

情感分析(sentiment analysis)，又称倾向性分析、评论挖掘(review mining)、情感挖掘或主观分析，是用户对商品、服务等评论内容的分析、处理、归纳和推理，对评论中表达的观点和情感进行分类，主要包括情感表达的主体、客体和内容。情感分析的研究主要集中情感分析方法以其应用两大方面。

情感分析主要针对文本信息，目前的研究主要集中在针对不同领域的市民评论进行分类、有效性、预测等方面。在旅游业领域，郭宇、王晰巍等人使用情感分析的方法，通过同程旅游网中的评论信息，构建了情感分析的用户影响力模型并绘制了评论信息的情感雷达图、用户影响力的幂律图和情感词的标签云图：在网络社交领域，王伟军、黄英辉等人通过爬取微博评论信息，利用《同义词词林》及 word2vec 等工具构建新的情感字典，进一步通过分析公众情感对新产品市场进行了预测研究：在新闻领域，潘云仙等使用 JST 模型对新闻文本进行分析，避免了与情感无关的语句对分析的影响。综上可得出情感分析对数据挖掘有一定的影

响。

情感分析自从 2002 年由 BoPang 提出之后，获得了很大程度的关注，特别是在在线评论的情感倾向性分析上获得了很大的发展。文本情感分类在情感分析研究中占有举足轻重的地位，在信息爆炸的 21 世纪，海量数据的情感分类研究吸引了很多的研究者，如何深入学习文本的语义信息，准确表达语义特征，提高情感分类的准确性是研究的目标。

目前，情感分析的主要研究方法还是一些基于机器学习的传统算法，例如，SVM、信息熵、CRF 等，机器学习的第一次浪潮是浅层学习，深度学习则是机器学习的第二次发展浪潮。以往的情感分析主要是采用浅层学习，但是无法学习文本语义信息，随着技术的发展和科技的进步，人们的要求也随之越来越高。在大数据的分析和处理上浅层学习存在的弊端导致情感分析遇到了瓶颈，因此人们将焦点转移到了可以改善这一弊端的深度学习的研究。目前对在线评论主要集中在数据挖掘，可视化，文本分类等方面，相关学者已经取得了较多研究成果。早在上个世纪,就已经有学者对产品评论进行对消费者影响的研究。随着互联网的发展，人们更倾向于在网络平台上发表对产品评价，在线评论已经成为大众评论的主要形式。周纯洁等通过机器学习的方式，分析了网络在线评论文本中网民的立场和观点信息：卢伟聪等对手机市场的评论进行分析，并结合二分网络分析探究了产品特征与用户评论之间的联系与用户的情感倾向。

但不是所有的用户评论都是有效的，只有含有有效信息并得到浏览用户认可的评论才具有实际价值。G.Ipeirotis 和 A.Ghose 通过研究评论中的主观和客观成分，分析其对用户评论有效性的影响；国内学者郝媛媛等基于用户评论数建立了评论有效性模型，并进一步对评论有用性进行了预测；刘志明基于说服双过程模型通过 IMDB 和豆瓣网的影评数据分析了在跨文化视角下的评论有效性。上述研究通过研究评论本身及其相关因素，探究了用户评论的实际价值，但是并没有讨论打分评价与评论之间的关系。马松岳、许鑫，使用 ROSTEA 工具进行情感分析得到评论评价的综合情绪值，将其与打分评价进行相关分析；M.E.Basiri 等基于心理发现和消极偏见的理论，利用评论历史改善情感分析，在细粒度的评论内容层面上通过评论等级评分来预测整体等级。这些研究表明打分评价与评论之间有着紧密的联系，所以本文尝试使用情感分析的方法，建立新的打分评价模型，与原有的打分评价进行对比，分析二者的差异性与相似性。

②这里我们结合基于数据和情感分析的用户评价建立模型，该模型分为 5 个模块，包括数据获取与预处理模块、数据库构建模块、情感分析模块、二分网络模型以及点赞数与反对数对比模块。

由于题目已给出详细数据，接下来我们直接通过情感分析模块调用短评文本对语句进行情感分析，并给出单句的情感得分。本文使用 BosonNLP 情感字典作为词汇本体，对其进行词表扩充。同时使用否定词词典、程度副词词典和停用词词典作为情感分析的基础计算单句情感得分的计算，如公式(1)

$$senti\_score = (-1)^n \sum\sum v_{ij} w_i sensibility_i \qquad (1)$$

其中，$sensibility_i$ 代表第 i 个情感词，w_i 代表第 i 个情感词的权重，$(-1)^n$ 代表否定词的额个数，v_{ij} 代表 i 个情感词的第 j 个程度副词。可以看出，由于程度副词取值均为正，所以情感得分的正负一方面取决于情感词，如果是积极情感词，那么 $sensibility_i > 0$，如果为消极情感词，那么 $sensibility_i < 0$，而当中性词 $sensibility_i$ 的接近于 0；另一方面则取决于否定词的个数。由于每条评论的字数与内容没有

进行限制.所以该计算方法的取值范围为正无穷到负无穷,即如果评论的正面评价越多则得分越高，负面评价越多则得分越低。所以为使得方便于下文的讨论，将得分进行归一化处理，并保证所有得分在 0-5 分之间，如公式(2)

$$senti\_norm = 5*\left[senti\_score + \min\left(senti\_score\right)\right] / \max\left(senti\_score\right) \quad (2)$$

其中 $\min\left(senti\_score\right)$ 代表所有得分中的最小值， $\max\left(senti\_score\right)$ 代表所以得分的最大值。并将所有得分划分为 0-1.5 分、1.5-3.5 分和 3.5-5 分三个区间，分别表示消极情感评论，中性情感评论以及积极情感评论。进一步在二分网络模块中，本文结合复杂网络分析，构建用户评论与评论特征的二分网络，并通过网络可视化处理，分析二者之间的联系以及评论用户的情感倾向。由于不同评论者的评论内容不同，所以每条评论信息对评论特征也不相同，通过构架用户评论与评论特征的单顶点二分网络，分析评论特征对政府工作认可度的影响，进而对不同的评论设置不同的权重，综合单句情感得到评分，如公式(3)

$$film\_senti = \sum\sum f_{ij} senti\_norm_i \quad (3)$$

其中， f_{ij} 代表第 i 个单句情感得分的第 j 个影评特征的权重。

由于电影情感评价评分只考虑的了评论者的短评文本信息,而忽略了其评分信息,所以进一步通过评论者的评分信息计算电影得分，即用户评价评分，计算公式(4)如下 $star\_score = \sum star\_p_i \times star_i$ （4）

其中，$star_i$ 代表的用户对电影的评分，分为 1-5 分五个整数分数；$star\_p_i$ 代表了分数 i 占所有打分人数的百分比。

得分与对比模块

③最后在得分与比较模块中，该模块主要基于上述的情感分析与网络分析，通过得分的形式将分析结果具体表现出来，对实际研究对象进行阐述。不论是对政府工作认可度的总分，还是对政府工作的相关意见均使用五分制，本文并假定两者在得分占有同样的地位，将两个得分直接相加得到最终评分，并将最终得分与最初计划的实施结果进行比较,分析原始与本文建立的评分模型之间的联系与差异,得到情感角度下的多维度影评信息。

2.2.5 步骤五：答复意见分析

随着经济和社会的快速发展，社会公众的物质需求得到了极大满足，更高层次的精神文化需求成为社会公众的奋斗目标。如何有效地满足社会公众的精神文化需求，成为亟待解决的问题。2014 年 12 月 25 日，十二届全国人大常委会第二十五次会议审议通过了《中华人民共和国公共文化服务保障法》，该法明确公共文化服务应结合当地实际需求和公众意见，在公共财政支撑能力的基础上充分挖掘和利用地方文化资源，提供人民群众喜闻乐见的公共文化服务。该法的颁布为公共文化服务提供了法律保障，但是政府如何挖掘和利用地方文化资源，社会公众如何表达自身实际需求，并没有得到有效解决。社区是现代社会公民的生活共同体，也是政府提供公共文化服务的主要载体。本文立足城市社区，以“解剖雀”的方式分析现阶段社区公共文化服务表达存在的问题，分析原因，并提出相应的对策。

一、社区居民公共文化服务需求表达的困境

笔者采用田野调查的方法，实地走访了一些社区，与居委会主任等一线工作人员进行了深入座谈，笔者发现现阶段社区居民在表达需求存在不愿表达、无渠道表达、表达得不到反馈等问题。（一）社区居民公共文化服务需求表达的主体缺位。1.社区居民缺乏需求表达的主动意识。以《海南省基本公共文化服务标准

化均等化先行区建设实施方案》为例，海南省人民政府办公厅制定了详细的公共文化设施项目建设方案，并且规定了详细的时间节点。但是这整个过程中并没有社区居民的参与，社区居民缺乏表达自身公共文化服务需求的主动意识。主动意识是居民表达的内生动力，内生动力不足导致社区居民愈加忽视自己对于公共文化服务的知情权、选择权和监督权，忽视了自己的文化权利。2.代表社区居民利益的社会组织的作用不够。根据政治学的相关研究，需求表达的有效性和表达主体的组织化程度呈正相关关系，组织化程度越高，表达的力度就越大，效果越明显，对政府决策的影响就越大；反之，需求表达的力度和有效性越小。我国的居民社区经历了从“单位”时代到“后单位时代”的转变，“单位”时代的社区作为单位的附属，具有很强的组织性，社区居民的很多需求可以通过单位的机构获得充分的表达，单位的机构也能够代表社区居民向政府表达相应的需求。“后单位时代”的社区组织化程度降低，社区居民因兴趣爱好成立的松散组织无法代表社区居民向政府有效表达相应的需求。（二）社区居民欠明确。表达渠道不畅，导致政府与民众之间缺乏连接的桥梁。其中最典型的表现就是居委会作为基层群众性自治组织，并没有充分发挥作用。根据笔者的调研，现阶段的居委会主要是协助基层政府完成相关的数据统计、信息收集、文件宣传等工作，换句话说，居委会在履行“下达”的职能。但是关于社区居民的需求是什么则较少收集，也很少及时向上反馈。当然必须说明的是，造成现状的原因是多方面的。（三）社区居民的需求没有得到及时有效地回应。当社区居民向政府及相关部门表达需求时，不管政府是否采纳，都希望得到及时的回应和反馈。意见被采纳固然好，不被采纳，能够得到政府及其职能部门的反馈并说明原因，也能够增进政府和社区居民之间的相互理解。这是参与型政府的基本内涵。但是根据笔者的调研，现阶段政府对于社区居民的需求很好回应。社区居民的需求长时间大面积得不到及时的回应势必会挫伤社区居民表达需求的积极性。

二、社区居民公共文化服务需求表达不畅的原因分析

前已述及，社区是构成社会的基本单元，是文化建设的基础。之所以社区居民的需求表达存在以上困境，笔者拟从社区和政府两个视角进行观察，找寻原因。（一）社区层面不同于乡村，“后单位时代”的城市社区人口密度大、异质化程度高、流动性强，对于公共文化的需求，表达的意愿都各有不同。让每一个社区居民都能表达自身需求显然并不现实。但如何有效组织社区居民，通过相应的制度设计和工作流程，既保障社区居民的表达权，又过滤掉社区居民的不合理的需求，着实是个难题。（二）政府层面实践中，政府往往根据自身的人力、物力、财力等情况通过推理计算形成理性的公共文化服务供给规划，这种理性规划的假定前提是人们的文化需求具有同质性。但就单个居民而言，不同的文化背景、社会阶层、经济地位及环境资源等都会影响他们的意识和需求，个体的需求具有差异性。也就是说，政府并没有围绕保障社区居民的文化权利来安排自身的行为。相较于社区，政府一方面应确定提供相应的文化服务是政府的义务，另一方面应进行相应的制度设计保障社区居民的参与权。让社区居民参与到公共文化建设中，让政府有限的财力发挥更大的作用。但很遗憾，政府并没有按照这个方向有效地发挥作用。除此之外，很多社区居民对表达非常不自信，认为政府能作出远优于他们的判断，政府不需要他们的建言。再者，中国的传统文化比较含蓄，表达自身需求总被认为是一件不好的事情。传统的影响、心理的不自信、思维的惯性等综合因素叠加导致社区居民不愿表达。观念的变革是制度生成的先导，要想真正改变这种现状，必须首先改变社区居民普遍存在的观念。通过宣传手段明确社区

居民的主体地位，明确社区居民的主体地位，提升社区居民的表达意识。政府及其职能部门通过收集社区居民的诉求，安排社区公共文化的建设，逐渐形成社区居民和政府职能部门的良性互动机制。（二）借助现代技术工具，拓宽社区居民的表达渠道。只有广泛听取公众的意见，政府的决策才能科学有效，才能易于被公众所接受。针对社区居民表达渠道不畅的问题，笔者认为应通过现代技术工具拓展表达渠道。具体来说，可以通过微信、微博、论坛等社交平台发布征集公共文化服务建设的信息，社区居民亦可通过以上渠道表达自身需求，政府通过后台收集并汇总相关的信息。3 月 26 日，全国第一个省级“互联网+”公共文化平台——“文化上海云”正式上线运营。“文化上海云”整合了海量的文化资源和用户，通过大数据分析，可以挖掘出公民最需要的公共文化服务信息，据此做出公共文化服务决策并组织相应的供给，使公共文化服务有的放矢，提升公共文化服务效能。“文化上海云”是互联网时代构建现代公共文化服务体系的样板，代表了未来公共文化服务建设的基本方向。海南可以充分借鉴上海的有益经验着力打造海南文化服务平台。除此之外，未来还可以在公共文化服务领域引入人工智能系统，通过人工智能系统收集、汇总、分析、反馈各种信息，达到公共文化的精准服务。（三）按照参与型政府的建设要求，明确政府的反馈义务。政府对社区居民的意见反馈义务是指在社区居民表达公共服务需求的过程中，政府针对社区居民提出的意见和建议进行收集、整理和研究，并针对采纳情况和具体理由作说明回应的义务。从现有立法看，有关规定明确了社区居民“进入表达程序”以及“提出意见”的权利，但在对“意见得到回应的权利”的规定上则明显不足。显然，社区居民的参与权出现了权能断层。在政府欠缺反馈义务来源的情况下，社区居民实现参与权存在法律上的障碍。同时，政府在观念上对社区居民表达的排斥，也进一步加剧了对社区居民参与反馈的忽视。笔者认为，反馈是建立互动机制的不二法门，必须明确政府的反馈义务，才能助推社区居民表达权的行使。具体来说，反馈活动应当遵守一定的原则。

与此同时，理顺政府与社区居民两者在反馈活动中的权利义务关系。参与权是相对人在行政程序中的一项重要权利，从行政主体之间的权利与义务的平衡来看，社区居民提出意见或建议的权利应当与政府对社区居民的回应义务相对应。倘若欠缺了回应便使得双方权利义务失去均衡。

三、中国现状并进行分析

近十年来，中国网民数量不断在增长，已经成为中国公民社会中最庞大的群体，与此同时，社会转型产生、积累的矛盾不断爆发，民众的民主意识不断觉醒，对个体利益的关注度越来越高，他们大量地在网络上发布诉求，迫切地希望获得政府的回应，而各级政府也注意到了这种庞大的需求，越来越关注网络言论，重视网络舆情，并通过各种渠道作出回应和引导。网络问政就是在这样的背景下应运而生，成为有别于传统媒介的政民沟通的新渠道，并蓬勃发展起来。

随着网络问政普及率越来越高，其自身的缺陷和发展的瓶颈也渐渐突显，从问政的环境到问政的主体再到问政的平台，都不同程度地陷入困境。政府在网络问政中的行为主要体现为政府回应，这也是民众在网络问政中最期待获取的元素。政府回应做得好，网络问政取得效果、实现价值，才能可持续发展；反之，回应做不好，民众不来问，网络问政也就失去了存在的意义。基于对网络问政实际运作的观察、梳理与认知，本文认为“政民互问”定义下的网络问政是理想模式，网络问政应当是一个政府和民众相互要求、相互影响、相互促进的过程。但是，鉴于中国目前仍处于民主的初级阶段，民众对政府的问责远多于政府对民众的问

计，政府被动回应远多于政府主动回应和问计，目前网络问政更多的是“民问政”。因此本文认为，现阶段下网络问政是民众通过网络媒介，表达利益诉求，行使民主权利，收获政府回应的过程。

为解决上述问题,全面及时准确了解企业和群众对政务服务的感受和诉求,接受社会监督,有针对性地改进政务服务,提升政府工作效能,优化营商环境,建设人民满意的服务型政府,问政效果的主要体现就是政府回应的满意度高低，本文对问政平台上的政府同应进行满意度分析，选择“已回复”状态下，参与了满意度评价的帖子，对其进行量化和质化的分析。量化的目的是为了通过数据归纳考究政府回应在哪方面对满意度评价产生了影响，质化的目的是为了通过实证分析不同满意度的政府回应的特点和存在问题。政府回应时效性较高，态度普遍明确，让民众体会到了当地政府部门的关注和重视。一些高质量的回应也体现了政府为民服务的立场和决心，展现了政府部门专业的形象，对提升政府公信力有积极的影响。与此同时，满意度评价的对外开放，表明了当地政府愿意接受群众监督和评议的态度。可见，“零距离问政”在促进政府和民众沟通互动、引导民众发挥监督评议作用、帮助民众解决诉求这三大方面，都取得了一定的成效。

以北京为例,在各级政务服务场所的办事窗口设置评价器或评价二维码,方便企业和群众自主评价,实现现场服务“一次一评”;在各类政务服务平台设置办事评价环节和评价功能模块,方便企业和群众即时评价,实现网上服务“一事一评”,没有及时作出线上或线下评价的企业和群众,可在 5 个工作日内进行补充评价。评价方式有:现场服务“一次一评”、网上服务“一事一评”、社会各界“综合点评”、政府部门“监督查评”。评价一般设置“很好”、“好”、“一般”、“差”、“很差”或“非常满意”、“满意”、“基本满意”、“不满意”、“非常不满意”五个等级,后两个等级为差评。

双向互动原则。“网络问政”是个双向互动过程。网络问政的主角是各级领导干部和广大网民，应建立一个上下衔接的工作机制，使“网络问政”步入制度化、规范化轨道，形成网络信息督办联动机制，真正做到问政于民、问计于民、纳智于民。在突发舆情危机应对和网络舆论引导中，需要深入和网民互动交流，尤其是活跃和知名的网友，一方面是集纳民意，一方面是沟通引导，说明存在的困难和问题，对于偏激的观点，要进行平等、公开的辩论，保持冷静和谦抑性，坚持文明上网，注意网络言行，体现良好的公共素质。坚持理性和建设性，实现双赢。所谓理性，就是理智地发表言论、控制行为的能力，亦即在面临某种问题时，能够注重思考、讲究方法，尽量冷静审慎地发表意见和采取行动。而建设性是以积极、正面的心态去看待事物、解决问题，其核心在于利益相关方之间的相互理解、尊重及未来导向。一般来说，理性主要是指态度，建设性主要是指方式方法，二者是统一的。所谓“双赢”，显然是理性、建设性地处理人民内部矛盾的最佳选择。

3、结论

网络论坛是社情民意调查、分析与监测最重要的数据来源之一，因此高效、准确获取论坛网页数据是社情民意调查、分析与监测的重要基础。传统的方法是针对每一种论坛类型做一个相应模板来获取论坛数据，然而网络论坛数据格式的多样化，必然使得系统的工作量巨大，而且可扩展性差。为了解决这一问题，本论文提出了一种基于 VIPS 算法和坐标信息的论坛数据记录提取算法,克服了传统网页数据提取方法的缺点，在保持了较高准确性的前提下，大大提高了系统的可扩展性和通用性，效果良好。

社区公共文化服务建设是公共文化服务建设的重要内容，通过田野调查等方式，解剖麻雀，以小见大，能够明晰现阶段社区公共文化建设社区居民表达方面存在的问题，社区居民不愿表达、没有渠道表达、表达得不到及时回应。直面这些问题，找寻背后的原因，社区居民公共文化服务需求表达机制的建构。一是提高社区居民的权利意识，二是应借助现代科技搭建更便捷的表达渠道，三是明确政府的反馈义务，多措并举，构建社区居民公共文化服务的需求表达机制。

4、参考文献

[1]王敏. Web 文本聚类算法在基于竞争情报的智能决策支持系统中的应用[D]J.北京科技大学硕士学位论文，2005

[2]Deng Cail,Shipeng Yu,Ji-Rong Wen,Wei-Ying Ma. Block-based Web Search SIGIR'04,July 25-29, 2004, Sheffield, South Yorkshire, UK

[3]俞士汶.计算机语言学概论[M].北京：商务印书馆 2003.

[4]Deng Cail,Shipeng Yu2,Ji-Rong Wen,Wei-Ying Ma. Microsoft Technical Re-port,MSR-TR-2003-79,2003

[5]Chakrabarti S.,Punera K and Subramanyam M.,Accelerated focused crawling through online relevance feedback,In Proceedings of the eleventh international conference on World Wide Web(WWW2002),2002, pp.148-159.

[6]Chen J,Zhou B, Shi J,Zhang H-J,and Wu Q,Function-Based Object Model Towards Website Adaptation,In Proceedings of the 10th International World Wide Web Conference,2001.

[7]Tang, Y. Y,Cheriet, M.,Liu, J.,Said, J.N.,and Suen, C. Y., Document Analysis and Recognition by Computers, Handbook of Pattern Recognition and Computer Vision,edited by C.H.Chen,L.F.Pau,and P.S.P.Wang World Scientific Publishing Company,1999.

[8]彭鹏.网络舆论的功能及调控策略[J].南京政治学院学报,2005(03)：115-116.

[9]王元睿.浅析微博舆论的特点、趋势及应对方法[EB/OL].[2011-07-22. http://www.yunxian.cn/Article/ShowArticle.asp? ArticlelD=33761.

[10]李丹.公民社会视角下中国微博舆情的发展与走向[J].东南传播,2011(5):6-8.

[11]达闻微指数[FB/OL].[2011-09-10].http://weibo.com/app/detail/6FiHzb 2011.9.1

[12]微博小分析[EB/OL].[2011-09-10].http://weibo.com/app/detail/6m3Wrd.

[13]社会网络分析方法和 IWOM 研究的结合初探[EB/OL].[2011-07-22]. http://www.seeisee.com/index.php/2009/07/22/p1350. I

[14] 盛宇，刘俊熙，郭金兰等.自然语言理解心理学在短文本分类中的实证研究[J].现代情报，2009,29(8):4-7.

[15]余波.微博的情报学意义探讨 J].图书情报工作、2010,54（22）:57-60.

[16]盛宇.基于微博的学科热点发现、追踪与分析—以数据挖掘领域为例.

[17]王强强，刘咏梅.《社区居民公共文化服务需求表达机制建构》

[18]卢伟聪,徐健.基于二分网络的网络用户评论情感分析[J].情报理论与实践 2017，41（2）:121-126.

[19] 赵妍妍，秦兵，刘挺.文本情感分析[J].软件学报,2010,21(8)：1834-1848.

[20]夏启政，董益好.基于情感分析的影评数据挖掘.2019.03.27.