文章编号:1671-4598(2018)03-0183-04 DOI:10.16526/j.cnki.11—4762/tp.2018.03.046 中图分类号:TP391 文献标识码:A

基于关注关系的互联网云数据挖掘方法实现

吕春荣，叶施仁
（常州大学 信息科学与工程学院，江苏 常州 213164）

摘要：随着互联网技术的飞速发展，互联网上的数据的获取也变得越来越简单与便捷；针对海量互联网云数据的数据挖掘技术的研究也成为网络数据研究领域不可缺少的一部分；然而由于现在互联网云数据的海量性，如何精准有效地从其中获取数据变得尤为重要；因此，提出了将数据网页以及内容间的引用与被引用行为抽象为一种关注关系的方法，根据对此关系的分析与综合处理，设计并实现了一种互联网云数据挖掘方法。实验结果表明该方法能够较好地实现对于互联网云数据的精准挖掘。

关键词：互联网；云数据；关注关系；数据挖掘

Implementation of Internet Cloud Data Mining Method Based on Following Relationship

Lü Chunrong, Ye Shiren
(School of Information Science & Engineering, Changzhou University, Changzhou 213164)

Abstract: With the rapid development of internet technology, it is more simple and convenient to get the data on the internet. So in view of the efficient mining of massive internet cloud data has become an indispensable part of network data field. Because of the mass of internet cloud data now, it becomes particularly important to get data accurately and efficiently from it. Therefore, this paper abstracts the reference relationship between the sites and content as the following relationship. Based on the analysis and synthesis of the relationship, an internet cloud data mining method is designed and implemented. The experimental results show that this method can achieve higher accuracy to the mining on the internet.

Keywords: internet; cloud data; following relationship; data mining

0 引言

作为当今最大的数据汇集地，互联网上的信息数据能够为用户带来极大的价值与效益，因此目前互联网是各项研究的热门[1]。同时由于互联网数据的海量性与各种广告数据节点的充斥，如何精准地从庞大的互联网云数据中获取到所需的信息资源成为了一项重要的研究[2-3]。

在互联网上根据需求搜索信息，用户普遍无法迅速、精准地回去到所需的有价值的信息对应的数据节点[4]。而从互联网中挖掘目标信息的方式又可分为内容挖掘、结构挖掘等[5]。其中内容挖掘通常是利用聚类分析、关联规则等方法获取所需的信息内容，以此为基础的挖掘方法均会存在匹配率较低以及冗余等不足[6]。

因此，考虑到互联网中的云数据以及网页之间存在的链接结构，这种数据节点与内容间的指向与被指向的关系，与现实生活以及数据结合点中的人与人之间的关系较为接近，因此本文利用分析数据结合点[7-8]的方式来进行互联网云数据结构进行分析。本文将互联网云数据结构的指向与被指向抽象为数据结合点中的关注关系，并以此为基础设计并实现了挖掘方法，能够较好地实现针对互联网云数据集数据节点的有效挖掘。

与一般的综合进行数据节点内容主题计算获取不同的是，本文提出的方案仅基于互联网数据节点间的关注关系来进行所需互联网数据的挖掘。本文基于数据节点间的关注关系定义了两个特征：关联度和影响力。定义关联度用来衡量待判断数据节点与已有数据节点集合间的连通强度，并以此判断该数据节点是否为所需主题。而利用影响力可以使得实验向相对正确的方向扩展分析发现目标数据以及数据节点。

本文选定目标内容相关的若干个公认较为精准数据节点作为种子集，并以此为基础扩展搜索获取种子集数据节点的关注数据节点（即该数据节点中存在的数据节点链接）中满足条件的部分，并不断迭代扩大数据节点池规模直到得到以种子集为代表的所需内容的主题数据节点集。最终实验结果显示基于关注关系的关联度和影响力能较好地进行互联网云数据发现。

1 互联网云数据挖掘模型

互联网上一个数据节点上存在数个数据节点链接，即可抽象为这数个数据节点被该数据节点关注，而该数据节点关注了其他数个数据节点。因此数据节点之间的这种“关注与被关注”的关系构成了数据节点之间的有向边。根据内部群体同质性，在一定条件下，在相同群体内部的关注行为是同质的[8]。拥有相似主题的数据节点之间存在很多关注关系，这将导致这些数据节点之间的连通密度显著提高。关注关系数据的获取过程相对简单，同时也具有稳定性。因此，使用关注关系来进行互联网云数据发现较为灵活且迅速。为了衡量数据节点之间的连通密度，我们定义了关联度；同时为了衡量已有数据节点集的权威排序，我们定义了影响力。互联网数据节点关联网络可以使用关注与被关注关系的有向图来表示[9]。

有向图 $G=(V,E)$ 表示微博的完整互联网集，其中 V 表示互联网集中的数据节点集，而 E 表示其中的关注关系集；待发现的数据节点集有向图 $G'=(V',E')$ 表示 $G=(V,$

收稿日期:2017-08-11； 修回日期:2017-08-29。

作者简介:吕春荣(1993-),男,江苏如皋人,硕士研究生,主要从事数据挖掘方向的研究。

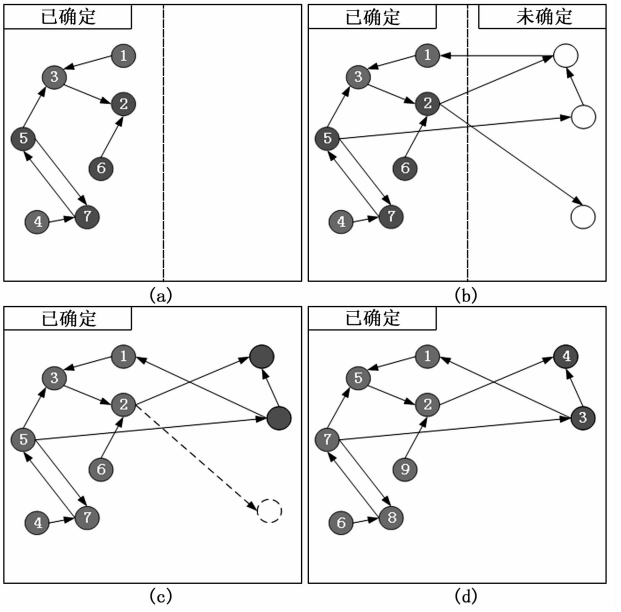

图 1 扩展过程简单示例

E）的一个子图，V’表示子图中的数据节点集，而 E’表示子图中的关注关系边集。并满足以下条件：

（1）V’中的所有点表示已判断的数据节点以及现阶段等待判断的未确定数据节点；

（2）E’中的所有边表示图 G’ =（V’，E’）中的所有数据节点的关注关系；

（3）图 G’ =（V’，E’）随着不断持续的数据节点挖掘发现在不断扩大。

有向图 G’ =（V’，E’）被构建来衡量计算一个未确定的数据节点与已有数据节点集之间的关联度。本文的所有实验工作均基于图 G’展开。

当本文的实验系统不断地扩展到新数据节点时，只会考虑那些关联度大于当前的关联度阈值的数据节点，而产生的不合格的数据节点将会在扩展过程中被丢弃以降低扩展的代价。扩展过程的简单示例如图 1 所示，图中有向箭头表示数据节点间的关注关系，浅色的点代表已确定并且已经过扩展的数据节点，深色的点代表已确定且正在进行扩展分析的数据节点，空心的点表示刚刚扩展获得且正在分析的数据节点，而虚线构成的点代表根据扩展修正算法而被舍弃的数据节点。

图 1（a）中左侧的点均为确定的数据节点，其中包括已扩展结束的浅色点以及尚未进行扩展的黑色点；图 1（b）中右侧空心点代表从深色点进行扩展得到的未确定点；图 1（c）表示根据右侧未确定点与左侧已确定节点团体之间的关联度进行的点取舍，其中被舍弃的点 的关联度不满足阈值；图 1（d）表示所有已确定节点进行影响力排序。

2 云数据关注关系分析

2.1 云数据关联度

直观地，某些数据节点由于主题的相关性，存在大量的指向与被指向关系即关注关系。因此本文基于某个数据节点和数据节点集中其他数据节点间的关注关系定义了关联度。此处利用数据节点的出度与入度来计算关联度[10]，低复杂度的计算能够提高实验系统判断响应速度且能降低针对大规模数据时的计算成本。简单来说，当某数据节点与目标数据节点集之间有更多的出度与入度，也就表明了该数据节点属于目标数据节点集的概率更大。但互联网中存在被指向达到数千万的数据节点，而一般的主题数据节点集中的数据节点同样也可能指向这些数据节点，导致数据节点集中有大量的出边指向这样的数据节点，这就使得这种数据节点在一般的主题数据节点集中也会拥有较高的关联度和影响力。这些不属于主题数据节点集的特殊数据节点如果加入进去，其海量的关注关系将导致互联网数据节点集发现偏离主题导致无限扩散而失败。为了避免这种情况出现，我们在关联度中引入了惩罚因子来抑制此类数据节点的加入。基于此得出关联度来判断新扩展数据节点是否属于图 G’。根据数据节点的出度与入度及惩罚因子定义关联度公式为：

$$C(p)=\frac{(w_1 In(p)+w_2 Out(p))}{In'(p)^2-Out'(p)^2} \tag{1}$$

其中：p 为等待判断的数据节点，In（p）为 p 数据节点的入度，Out（p）为 p 数据节点的出度，ω_1 与 ω_2 为入度与出度的权重，且 $\omega_1+\omega_2=1$，$\frac{1}{In'(p)^2-Out'(p)^2}$ 为惩罚因子，In’（p）为 p 数据节点的所有入度即有链接指向它的数据节点数，Out’（p）为点 p 的所有出度即该数据节点指向的数据节点数。

2.2 影响力分析

在互联网主题数据节点集中，若其中一个数据节点的多数指向数据节点与被指向数据节点属于该数据节点集，那么可以认为该数据节点在该数据节点集中拥有更高的影响力。为了计算出在等待扩展时的优先权，需要做的是衡量出图 G’中的所有数据节点的影响力。因为图 G’中的大部分数据节点均是活跃的，且具有相同主题的数据节点更趋于互相关注。所以基于节点之间的关注关系来衡量数据节点在数据节点集内的影响力。数据节点的影响力公式为：

$$F(p)=\omega_1\frac{In(p)}{N_{in}+1}+\omega_2\frac{Out(p)}{N_{out}+1} \tag{2}$$

其中：p 为等待判断影响力的数据节点，N_{in} 为图 G’中所有数据节点的入度之和，N_{out} 为图 G’中所有数据节点的出度之和，实际中的 N_{in} 值与 N_{out} 的值相等，In（p）为 p 数据节点的入度而 Out（p）为 p 数据节点的出度，ω_1 与 ω_2 为各属性的权重，$\omega_1+\omega_2=1$，它们值的调整象征着在计算中关注对象与用户对影响力的贡献比例。

3 云数据挖掘实验

本文选择关联度特征作为分析标准，并根据关联度设定加入云数据的阈值，计算阈值的公式为：

$$V=(1+\alpha)\frac{\sum_{i=1}^{N}C(p)}{N} \tag{3}$$

其中：C（p）即为图 G’内一数据节点与图 G’内其他数据节点的关联度值，N 为当前图 G’内数据节点数量，而 α 为一可变参数，它的范围为 -1.0 到 1.0，通过调整它即可调整产生的阈值 V 的高低，即可调整数据节点集的扩展梯度和最终的数据节点集大小。

本文还设定了一个步长参数，通过调整它可以控制每一轮扩展的规模。步长参数公式为：

$$d=N\times\beta \tag{4}$$

其中：N 为当前图 G’内数据节点数量，β 取值范围为 0

到 1.0，通过步长参数可以控制每轮扩展结果的步长。

算法步骤如下。

步骤 1：先选取一些某主题内公认权威数据节点作为种子集，这些种子数据节点构成了最初的图 G' 并标记为未扩展。

步骤 2：使用公式（2）更新图 G' 内已有数据节点的影响力值。

步骤 3：选取当前图 G' 中未扩展且影响力最大的数据节点，扩展获得其关注数据节点并将该数据节点标记为已扩展。

步骤 4：将一个新获得待判断的数据节点临时加入图 G' 得到 G''，使用公式（1）计算出他在图 G'' 中的关联度值 $C(p)$，若关联度值大于公式（3）计算出的阈值，则将该点加入图 G'，否则舍弃并重复步骤 3 和步骤 4 直到新找出的加入图 G' 的数据节点数量不小于公式（4）得到的当前步长。

步骤 5：重复步骤 2，步骤 3，步骤 4，直到目标数据节点集达到设定的规模大小或图 G' 内不存在待扩展的数据节点。

4　实验及分析

4.1　实验数据及结果

本文选取数据挖掘作为目标数据节点集的主题，并通过 2.4 节中的算法获取该互联网主题数据节点集。为了便于人工检测兴趣云数据发现方法的有效性，本文定义数据节点集的上限为 500，并通过调节各个参数进行组合共做了四组互联网主题数据节点数据挖掘发现实验，并对最终实验结果进行分析与比较。在表 1 中，可以看到各参数值以及最终数据节点集准确率。对于数据节点集中数据节点准确与否的判断标准是该数据节点的主题以及指向数据节点与所指向的数据节点主题决定。

表 1　实验各参数值及实验结果准确率

	式(1)与(2)中 ω_1 与 ω_2	式(3)中 α	式(4)中 β	准确率
Ⅰ	0.5 0.5	0	0.5	89.67%
Ⅱ	0.3 0.7	0	0.5	90.33%
Ⅲ	0.3 0.7	−0.2	0.5	82.67%
Ⅳ	0.3 0.7	0	0.2	90.67%

4.2　实验结果与分析

由表 1 的实验结果可以看出，调整参数所产生的 4 组实验中，本文的互联网主题数据节点集方法在数据节点集规模达到 500 的前期至少有 82.67%的准确率。尽管不同参数下数据节点集发现的准确率相差不大，但是实际得到的结果数据节点集中数据节点构成存在一些差异。

此处有必要对数据节点集中数据节点进行分类以便更好地进行展示及分析。与数据结合点中人员构成类似，最终数据节点集中数据节点可以分为 3 种：权威数据节点、稍重要数据节点以及一般数据节点，通常在互联网中中这 3 种数据节点会呈现出金字塔型结构，由于本实验属于互联网主题数据节点集的早期阶段，所以本实验的最终获得的数据节点集内数据节点分类情况见表 2。

表 2　实验发现的数据节点集内数据节点的各比例构成

	专家	普通从业者	业余爱好者
Ⅰ	46.33%	37%	6.33%
Ⅱ	38.33%	47.33%	4.67%
Ⅲ	8.33%	14.67%	59.66%
Ⅳ	46.67%	37.33%	6.33%

1）由于在计算影响力值时对用户数据节点的权重的提升，实验Ⅰ得到的云数据结果中有超过一半的用户分属于权威数据节点类别。然而实验Ⅲ中由于阈值门槛的降低，虽然在实验初期跟其他实验相比并无较大差别，但实验中的数据节点集中数据节点的准确率比其他降得都快，同时超过半数的数据节点都属于一般数据节点。

2）将实验Ⅳ与实验Ⅱ比较后，可以看到将扩展步长调到较小没有获得较大的正确率的提升，但扩展到相同的数据节点集规模却需要更多的轮次。同样将数据节点集规模扩展达到 500 规模，其他实验仅需 11 轮的同时实验Ⅳ却需要多达 21 轮扩展，这其中的扩展步骤将耗费更多的计算与时间。

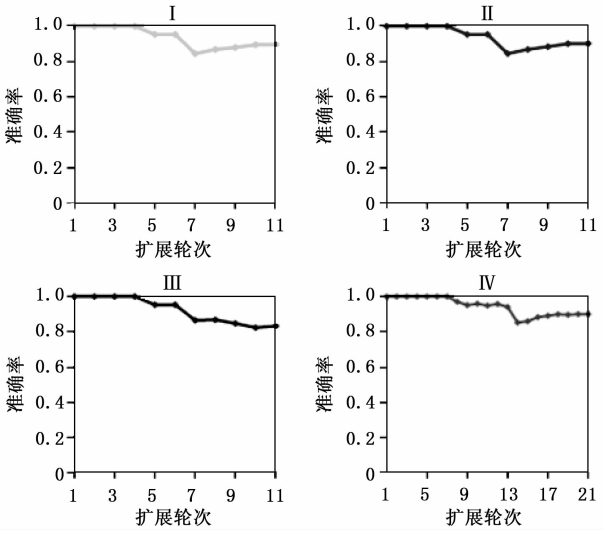

图 2　4 个实验在各扩展阶段云数据成员准确率

3）通过对最终数据节点集内错误划分到该集合内的数据节点分析，可以看到该错误数据节点会有较多的真正属于该数据节点集的用户指向。产生错误的原因是该错误数据节点在特定的热点主题中拥有较高的影响，例如购物、健康、股票以及热点主题，因此数据节点集内较多数据节点对该数据节点的指向关注导致了对该数据节点的错误划分。这种现象可能在一轮扩展里出现多次，而该轮导致的结果就是准确率会急剧下降。在图 3 中，实验Ⅱ的第 7 轮扩展由于错误划分了 12 位用户，导致了准确率出现了较大的下降。本轮结束后，根据实验步骤对已有数据节点集内数据节点的影响力进行计算并排序，可以有效抑制错误在下一轮中进一步扩大。因此第 8 轮扩展可以看出错误并没有延续，且最终正确率达到了一个较高的水平。可以看到，较弱的关联度阈值对于错误的控制并没有起到较好的效果。

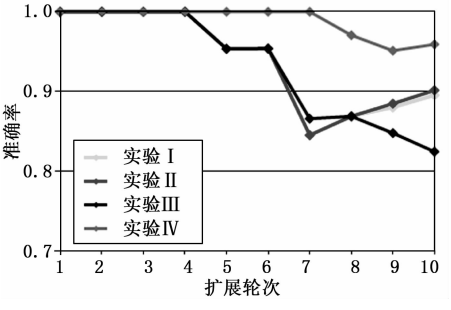

图 3　在前 10 轮产生的错误及对其控制的细节

5 结束语

基于互联网链接关系与数据结合点中关注关系的相似，将链接关系抽象为关注关系并基于此提出了一种互联网云数据的挖掘方法，并由此定义关联度和影响力作为特征来进行主题数据节点集的发现，且最终划分到数据节点集内的数据节点具有较高的正确率。本文的实验论证了提出的发现方法互联网云数据环境下发现主题数据节点集的有效性，且通过对过程中各项参数进行调整，可以调整最终数据节点集结果中的各种类型数据节点的比例构成。根据本文方法的特点，此方法最终可以实现准实时互联网主题云数据挖掘，可以更好地运用于针对互联网云数据的获取与使用的便利。在后继的研究中，我们计划继续挖掘互联网链接关系的潜在特性，以减少数据节点集发现过程中的错误产生和扩散，以期互联网云数据挖掘有更好的效果。

参考文献：

[1] 陈 琳，李 勇，王 磊. 面向移动互联网的不良信息监控系统设计 [J]. 计算机测量与控制，2016，24 (9)：126－129.

[2] 崔道江，陈 琳，李 勇. 智能检索引擎中的网络数据挖掘技术优化研究 [J]. 计算机测量与控制，2017，25 (6)：189－191.

[3] 林明方. 异构式分布下的 Internet 数据挖掘方法优化研究 [J]. 计算机测量与控制，2017，25 (7)：282－284，289.

[4] 乔智勇，刘志镜. Web 数据挖掘系统的设计及实现研究 [J]. 计算机工程与设计，2002，23 (7)：36－38.

[5] 曼丽春，朱 宏，杨全胜. Web 数据挖掘研究与探讨 [J]. 现代电子技术，2005，28 (8)：3－6.

[6] 李鑫洪，李庆华，刘炜娜. 国内 Web 数据挖掘研究综述 [J]. 现代计算机：普及版，2013 (12)：14－18.

[7] 邢东东，王秀文. 基于微博媒体的云数据发现技术研究 [J]. 智能计算机与应用，2013，3 (6)：74－77.

[8] 孙怡帆，李 赛. 基于相似度的微博社交网络的云数据发现方法 [J]. 计算机研究与发展，2014，51 (12)：2797－2807.

[9] 余永红，向晓军，高 阳，等. 面向服务的云数据挖掘引擎的研究 [J]. 计算机科学与探索，2012 (1)：112－132.

[10] 丁 静，杨善林，罗 贺，等. 云计算环境下的数据挖掘服务模式 [J]. 计算机科学，2012 (S1)：56－65.

[11] 邓仲华，刘伟伟，陆颖隽. 基于云计算的大数据挖掘内涵及解决方案研究 [J]. 情报理论与实践，2015 (7)：78－89.

[12] 朱亚东. 云计算网络中边界节点识别方法改进研究 [J]. 计算机测量与控制，2017 (1)：211－214.

[13] 张生福. 云计算虚拟现实技术供应链协同系统设计与实现 [J]. 计算机测量与控制，2017 (6)：98－112.

[14] 古忻艳. 网络计算机模型下海量大数据存储系统设计 [J]. 计算机测量与控制，2017 (6)：55－71.

[15] 何 清，庄福振，曾 立，等. PDMiner：基于云计算的并行分布式数据挖掘工具平台 [J]. 中国科学：信息科学，2014 (7)：88－117.

(上接第 182 页)

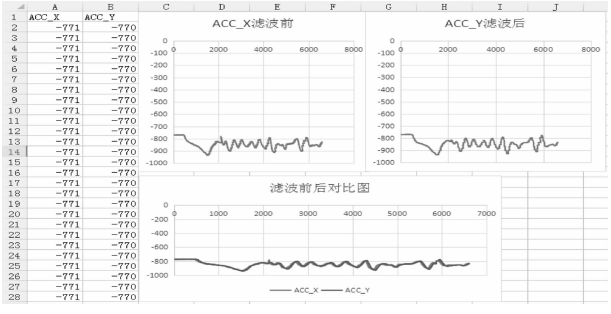

图 7 悬停相对中端位置数据滤波前后对比图

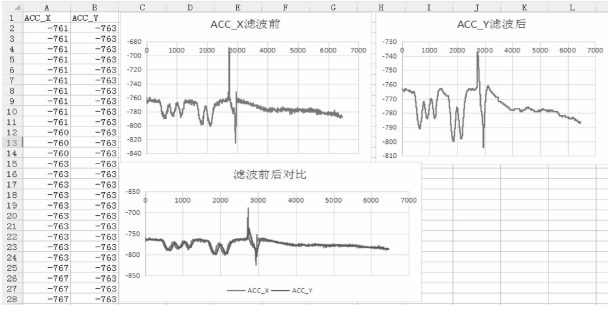

图 8 悬停相对低端位置数据滤波前后对比图

4 结论

悬停稳定性具有重大价值推广，因其实际飞控操作还可以运用到民用、商业场合，本文针对增加悬停稳定性的需要，通过以校园合作公司提供的良好开发平台，无人机丰富的数据采集，设计的扩展卡尔曼模型，软件使用，软件编程调试完成了阶段性无人机数据采集扩展卡尔曼滤波增加悬停稳定性研究与应用。在今后的无人机研究中将更进一步完善，运用更智能先进的算法提高无人机悬停运行稳定性。

参考文献：

[1] Plan Y，Vershynin R. One－bit compressed sensing－by－linear－programming [J]. Communication－on－Pure&Applied－Mathematics，2013，66 (8)：1275－1297.

[2] Hayashi K，Nagahara M，Tanaka T. A user's guide to compressed sensing for communications systems [J]. IEEE Transactions on Communications，2013，96 (3)：687－712.

[3] Blumensath T. Compressed sensing with nonlinear observations and related nonlinear optimization problems [J]. IEEE Transactions on Information Theory，2013，59 (6－7)：3466－3474.

[4] 李庆鑫，宗 群，王 芳，等. 基于鲁棒自适应的无人机直升机悬停控制 [J]. 控制工程，2014，2 (2)：253－257.

[5] Seassa S，Zecca M，LinZ，et al. A Methodology for the performance Evaluation of Inertial Measurement Units [J]. Journal of Intelligent&Robotic Systems，2013，71 (2)：143－157.

[6] 马正华，贺小棒. 基于预估测量值的 EKF 在手臂测资中的应用 [J]. 计算机测量与控制，2016，24 (11).

[7] 卢 浩，倪洪启，赵艳春，等. 无人机水下定点悬停平衡模块设计 [J]. 机械工程师，2017 (6)：31－33.

[8] 都基焱，张 振. 十字翼布局无人机悬停阶段控制律设计 [J]. 电子设计工程，2014，22 (3)：94－96.

[9] 王海洋，路 平，江 涛. 三旋翼构型倾转旋翼无人机建模与悬停控制研究 [J]. 电光与控制，2015，22 (10)：51－55.

[10] 王大鹏，王茂森，戴劲松，等. 四旋翼飞行器悬停建模及控制 [J]. 兵工自动化，2017 (5)：36 (5).

[11] 申文斌，裴海龙. 改进的 Unscented kalman 滤波算法 [J]. 计算机工程与科学，2011，33 (4)：143－157.

[12] Charles K. Chui，CHEN Guanrong. 卡尔曼滤波及其实时应用 (第四版) [M]. 北京：清华大学出版社，2013.

[13] 王世元，黄锦旺，谢智刚，等. 非线性卡尔曼滤波器 [M]. 北京：电子工业出版社，2015.