Open Journal of Transportation Technologies 交通技术, 2019, 8(2), 86-93
Published Online March 2019 in Hans. http://www.hanspub.org/journal/ojtt
https://doi.org/10.12677/ojtt.2019.82011

Research on Preprocessing Method for Vehicle License Plate Recognition Data

Wenhao Li[1,2], Hong Wan[2], Junhao Zhang[1], Chenxing You[1]

[1]School of Architecture and Transportation Engineering, Guilin University of Electronic Technology, Guilin Guangxi
[2]Maritime College, North Bay University, Qinzhou Guangxi
Email: 994714300@qq.com

Received: Jan. 31st, 2019; accepted: Feb. 14th, 2019; published: Feb. 21st, 2019

Abstract

In order to distinguish the effective use of license plate recognition data, this paper analyzes the characteristics of vehicle license plate recognition data and data anomaly, combining data preprocessing strategies in data mining technology to design a vehicle license plate recognition data. This preprocessing method includes four processes of data cleaning, data integration, data conversion, and data reduction. The effectiveness of the method is demonstrated by using the license plate recognition data of the detection bayonet in Guilin. The results show that the method can improve the utilization quality of license plate identification data effectively and reduce the impact of data anomaly on the model.

Keywords

License Plate Recognition Data, Data Mining, Preprocessing

面向车牌照识别数据的预处理方法研究

黎文皓[1,2]，万　红[2]，张钧浩[1]，尤晨星[1]

[1]桂林电子科技大学，建筑与交通工程学院，广西 桂林
[2]北部湾大学，海运学院，广西 钦州
Email: 994714300@qq.com

收稿日期：2019年1月31日；录用日期：2019年2月14日；发布日期：2019年2月21日

摘　要

为了实现对车牌照识别数据的有效利用，本文在分析车牌照识别数据特点和数据异常情况的基础上，结

文章引用: 黎文皓, 万红, 张钧浩, 尤晨星. 面向车牌照识别数据的预处理方法研究[J]. 交通技术, 2019, 8(2): 86-93.
DOI: 10.12677/ojtt.2019.82011

合数据挖掘技术中的数据预处理策略，设计了一种针对问题车牌照识别数据的预处理方法，该方法包括数据清理、数据集成、数据转换、数据归约4个过程。利用桂林市检测卡口的车牌照识别数据对该方法的有效性进行论证，结果表明该方法可以有效提升车牌照识别数据利用质量，降低数据异常对模型的影响。

关键词

车牌照识别数据，数据挖掘，预处理

1. 前言

车牌识别技术(Vehicle License Plate Recognition, VLPR)能够对检测路段通过的车辆自动识别并获取车辆牌照信息进行处理[1]。城市路网和高速公路网每天都在获取数以百万计的车牌照信息，与其它交通信息采集技术相比，车牌自动识别系统工作具有连续性强、数据精确度高、检测样本量大等优点。目前车牌照识别数据已经广泛应用于监测报警、超速违章处罚、车辆出入管理等方面，具有十分重要的现实意义和经济价值。

近年来国内外学者越来越多对车牌照识别数据进行深入挖掘，余丰茹等[2]通过车牌照识别数据，给出了适应于不同设计速度的高速公路交通拥挤识别方法。刘泉[3]探讨了车牌照识别数据提取交通流参数的基本方法，基于 J2EE 架构开发了交通流数据获取、处理和提取一体化数据共享平台。陈熙怡[4]在传统 OD 估计的基础上，提出了一种基于车牌照数据获取路网 OD 的方法。但目前关于车牌照识别数据的研究主要集中在车牌照数据的应用模型研究上，忽视了对车牌照数据预处理的研究。现有文献中鲜有涉及车牌照数据异常的研究，作为模型的基础，数据质量直接影响最终的建模效果，因此车牌照数据的预处理显得尤为重要，基于上述问题，本文针对车牌照识别数据海量性、多样性、实时性的特点，提出一种面向车牌照识别数据的预处理方法，包含数据清理、数据集成、数据转换、数据归约 4 种策略。

2. 数据描述

车牌照识别数据通过 VLPR 系统采集，车牌识别是现代智能交通系统中的重要组成部分之一，应用十分广泛。它以数字图像处理、模式识别、计算机视觉等技术为基础，对摄像机所拍摄的车辆图像或者视频序列进行分析，得到每一辆汽车唯一的车牌号码，从而完成识别过程。VLPR 系统一般由车辆到达触发模块、车牌捕获摄像头、识别系统处理器等部分组成[5]。数据获取过程可以描述如下：首先，当车辆通过时，会触发系统开始工作，识别系统在接收到信号后给摄像头发送图像采集信号，摄像头在车辆通过瞬间拍下带有车牌的车辆通过照片，接着识别系统终端会将此照片上传到系统处理器，在图片中将车牌提取出来，对车牌中的详细字符进行分割识别，给出车牌具体的识别结果，车牌识别数据的过程如下图 1 所示。VLPR 系统识别的数据构成除了车牌号码外，一般还应包括车牌号种类、车辆经过检测设备的时间、车辆类型、车道编号、地点编号、地点描述、行驶方向等内容。

3. 数据预处理方法设计

3.1. 相关概念

现实世界中数据大体上都是不完整，不一致的脏数据，无法直接进行数据挖掘，或挖掘结果差强人

意。为了提高数据挖掘的质量产生了数据预处理技术，数据预处理技术为进一步的数据分析做准备，其目的是将未加工的数据经过一系列步骤转化为适合分析的形式，主要处理策略包括数据清理、数据集成、数据转换、数据归约等过程。

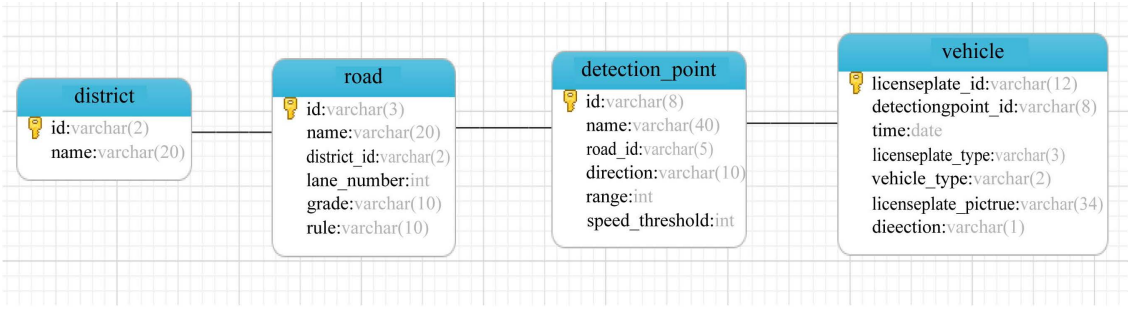

Figure 1. E-R relationship diagram of vehicle license plate identification data database
图 1. 车牌照识别数据数据库 E-R 关系图

1) 数据清理：处理原始数据中异常数据、重复数据、缺失数据等方法，包括数据分析、定义并执行清理规则、验证等步骤。

2) 数据集成：是把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中统一数据库储存的过程，主要解决数据的分布性和异构性的问题。

3) 数据变换：主要指对原始数据各属性进行规划化的相关操作，它涉及以下工作平滑、聚集、数据概化、规范化和特征构造。

4) 数据归约：是将数据仓库中海量数据进行整理归约，以较低的数据量保持数据完整性的过程。主要的处理策略包括维归约、数据压缩、数值归约和概念分层。其中，维归约(也可以称作特征选择)指的是删除与数据挖掘不相关的属性或合并相似属性，通常使用属性子集选择方法，找出最小属性集，使得数据概率分布尽可能接近原始数据分布，通过维归约能减少模式上的属性数目，使得模式更易于理解。数据压缩(也可以称作特征提取)通过对数据的压缩可以把数据储存在很小的空间中，在稀疏数据集的运算存储中应用较广，主要方法包括主成分分析、小波变换和多维标度法等。数值归约指利用替代数据以“较小的”数据表示形式来达到减小数据量的目的，其常用的方法有回归和对数线性模型、直方图、聚类及抽样等。概念分层是利用更高层次的概念替代低层次概念来定义数据属性的方法，概念化后的属性经过离散，其数据细节可能部分丢失，但其数据更有意义、更容易被理解。

3.2. 车牌照识别数据的预处理方法

在海量的原始车牌照识别数据中，由于调查点上的设备工作、人工操作、交通状况以及天气因素等原因的影响，很难确保数据的获取质量，往往存在冗杂数据、重复数据、缺失数据及错误数据等，车牌照识别数据主要异常情况如表 1 所示[6]。同时，在实际应用中，数据来自众多系统，具有多种形式和类型，其总是杂乱无章、不完全的，原有数据特征有时不能足够地体现隐藏在其后的规律，将严重影响到数据的实际应用。因此，有必要通过预处理来提高数据的“质量”。

根据车牌照识别数据自身特点来看，车牌照识别数据本身来源于不同位置的检测卡口设备，这样在每个调查点都会形成几个车牌照数据文件，除次之外，每天采集到的车牌照识别数据都是数以万计的，具有海量的数据规模，并且每条数据不仅是对车牌号的描述，同时包含车型、调查点、日期等信息具有多个维度，同时，由于检测设备的局限性，识别到的数据不可避免的会存在数据错误、数据漏检等问题。针对车牌照识别数据的特点和存在的问题，我们设计了一种基于 mysql 数据库和 C#的数据预处理方法，

该方法可以有效地对车牌照识别数据进行汇总、计算、校验等操作，具体方法如下：

Table 1. Abnormal situation of license plate identification data
表 1. 车牌识别数据的异常情况

主要异常情况	样例	归类
设备漏检	车辆经过时未被检测到	数据缺失
数据乱码	车牌号“&@!”	
检测错误	车牌号“桂 A281F4”识别成“桂 A281P5”	数据错误
数据逻辑关系混乱	车辆的时空关系不合理	

Step 1：数据集成

通常情况下车牌照识别系统可能涉及数十个甚至上百个信息采集点，就一个采集点而言，在正常交通状态下，平均每个车道每 3 秒通过一辆机动车，按 3 车道计算，每 1 秒钟便会产生两条记录。我们需要采用数据库技术对所有采集点的多种数据类型进行汇总和信息计算处理，数据库可以通过 SQL 语言很方便地对不符合规格的异常数据进行校验、清洗、补充，以最大限度地满足数据处理的基本需求。一条车牌照识别数据包括：序号、日期、车牌号、车型、调查点编号、交叉口、流向、通过时间、备注等信息，针对车牌照识别数据结构，本文设计的车牌照识别数据库由城区编码数据表、道路信息数据表、检测点数据表、检测车辆数据表 4 个表组成。图 1 为本文设计的车牌照识别数据库的 E-R 关系图；

Step 2：数据归约

车牌识别数据即所有车辆的检测记录集合，其包含多种属性，根据不同任务的需求，可以删除与任务不相关的属性或合并相似属性，即对车牌识别数据进行维归约。例如，基于车牌照识别数据的车辆出行轨迹挖掘算法中利用到的属性数据为车牌号、车牌号类型、检测时间、车辆类型、设备点位地址、设备点位编号、行驶方向及车道编号 8 项属性，其他的属性数据可作为“冗杂”数据剔除掉。同时，由于设备点位地址及设备点位编号均表征设备的空间信息，可进行合并，后期算法统一采用设备点位编号。

Step 3：数据清理

在车牌照识别数据采集过程中，受到设备识别率、信号传输干扰、套牌车等的影响，存在设备漏检、数据乱码、检测错误、数据逻辑关系混乱等问题。针对数据乱码、检测错误 2 种数据异常情况，传统处理方法会将车牌照识别数据舍去，但若问题车牌总量较多，舍弃问题车牌照识别数据将会对数据处理结果的准确性造成很大的影响。为充分利用问题车牌照识别数据，可基于调查点之间相邻关系和车辆途经调查点之间的时距关系，采用模糊匹配中的编辑距离算法来校核问题车牌。编辑距离是针对二个字符串的差异程度的量化量测，如 sun 到 son 编辑距离为 1，而 foot 到 feet 的编辑距离即为 2，编辑距离越小，相似度越高[7]。有如下计算公式：

$$d(i,j)=\begin{cases}0 & i=0 \text{ or } j=0\\ \min\left(d(i-1,j)+1, d(i,j-1)+1, d(i-1,j-1)\right) & x_i=y_i\\ \min\left(d(i-1,j)+1, d(i,j-1)+1, d(i-1,j-1)+1\right) & x_i\neq y_i\end{cases} \tag{1}$$

$$S=1-d(i,j)/L \tag{2}$$

式中：S 为相邻检测点之间的车牌相似度；$d(i,j)$ 为相邻检测点之间的车牌匹配编辑距离，其中 $d(i-1,j)+1$ 代表字符串 2 插入一个字母才与字符串 1 相同，$d(i,j-1)+1$ 代表字符串 1 删除一个字母才与字符串 2 相同，然后当 $x_i=y_i$ 时，不需要代价，所以和上一步 $d(i-1,j-1)$ 代价相同，否则+1，接着 $d(i,j)$ 是以上三者中最小的一项；L 为车牌识别字符串长度。

现有车牌识别技术达不到100%的精确识别，我们在进行车牌识别匹配时，需要设置一个阈值，只要两个车牌相似度在阈值范围内可视为同一车牌，车牌模糊匹配流程如图2所示。

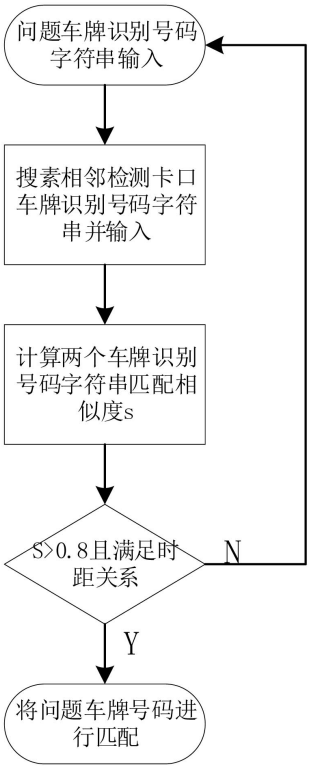

Figure 2. License plate fuzzy matching flow chart
图 2. 车牌模糊匹配流程图

数据乱码、检测错误2种数据异常情况在进行模糊匹配后，仍存在问题的数据，采取删除操作，删除方法考虑车牌号规则约束，车牌号约束条件包括：1) 常规车牌号仅允许以汉字开头，后面可录入六个字符，由大写英文字母和阿拉伯数字组成。2) 最后一个为汉字的车牌允许以汉字开头，后面可录入六个字符，前五位字符，由大写英文字母和阿拉伯数字组成，而最后一个字符为汉字，汉字包括“挂”、“学”、“警”、“港”、“澳”。3) 新军车牌以两位为大写英文字母开头，后面以5位阿拉伯数字组成。

而设备漏检和数据逻辑关系混乱2种异常数据需要借助车辆的路径数据进行识别，由于缺乏其他有效的信息对其进行还原，故本文的预处理主要针对数据乱码、检测错误2种数据异。

Step 4：数据变换

车牌识别数据中车牌号、设备点位地址、设备点位编号的字符串均比较长，占用储存空间比较大，我们采用数据编码的方式来表示车牌号、设备点位地址、设备点位编号的字符串达到减少数据量的目的。例如一个三位数编码000~999，唯一并简洁标识1000个不同条目，明显比每一条用语言描述占用空间少。

4. 实例

本文采用广西桂林市的车牌检测卡口数据，桂林市共设置了75套车牌识别检测卡口设备，图3展示了秀峰区车牌检测卡口所在位置，从地图中可以看到这些设备覆盖了桂林市的大部分主干道。数据库每天接收到几百万条记录，本文从中选择了两个数据集包含两个时间周期，第一个从2016年10月10日00:00:00到2016年11月1日23:59:59，第二个从2017年3月1日00:00:00到2017年3月21日23:59:59。

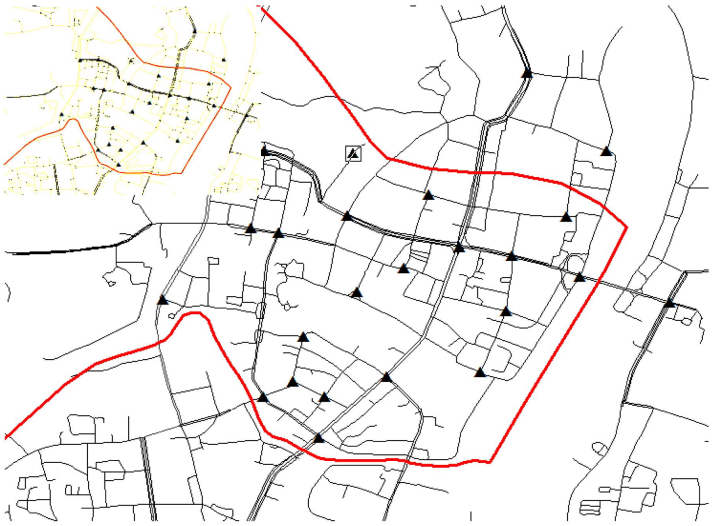

Figure 3. Detection bayonet position distribution map
图 3. 检测卡口位置分布图

两个时间周期中有个 6 双休日，其车辆出行规律与工作日会有明显差异，故本文删除这 6 个双休日的所有记录。仅使用剩余的 32 天工作日数据。

本文使用 mysql 数据库对车牌照识别数据进行管理，首先把所有的数据进行合并存入表 plate 中。在进行数据合并时，由于在数据录入过程中，每隔几分钟记录一次车辆通过时间并默认车辆在这几分钟内的到达是均匀分布的。因此，应先自动生成各个文件下全部车辆的通过时间，再补充车辆的序号、日期、调查点编号、流向等统一的信息，最后再将各数据文件按照时间顺序合并成一个文件[8]。然后从原始 26 个属性列中选择有用的列，可以使用 sql 语句“ALTER TABLE plate DROP COLUMN field_name”删除无用的列，合并后的数据格式如表 2 所示。

Table 2. Data set format
表 2. 数据集格式

列名	含义	示例
licenseplate_id	车牌号码	桂 CCCCCC
licenseplate_type	车牌号种类(01/02，01 为大型汽车；02 为小型汽车)	01
time	检测时间	2016-10-9-12:17:13
vehicle_type	车辆类型(0/1/2)	1
lane_id	车道编号(1，2，3，…从左侧车道数起)	2
detection_id	检测地点编号(从 1 到 75)	64
detection_name	检测地点名称	自由路与穿山路
direction	行驶方向	东向西
direction_id	行驶方向编号(01/02/03/04，01 为东向西；02 为西向东；03 为南向北，04 为北向南)	01

由于图像识别技术的局限性和数据传输过程中可能存在丢失数据问题、数据混乱等问题，我们需要对不符合规格的异常数据进行校验、清洗、补充，数据校验时首先根据预定的数据格式要求，对调查点的编号、车型、时间和日期进行校验和修正，然后再对车牌照的格式进行校验[8]。对于车牌号格式的校验，我们使用 C#连接 mysql，通过针对车牌号规则设计的正则表达式进行识别：

```
public static bool IsVehicleNumber(string vehicleNumber)
{
    bool result = false;
    if (vehicleNumber.Length == 7){
      string express = @"^[京津沪渝冀豫云辽黑湘皖鲁新苏浙赣鄂桂甘晋蒙陕吉闽贵粤青藏川宁琼
      使领 A-Z]{1}[A-Z][7][A-Z0-9]{4}[A-Z0-9 挂学警港澳]{1}$";
      result = Regex.IsMatch(vehicleNumber, express);
    }
    return result;
}
```

根据车牌照的格式校验结果，将问题车牌照识别数据按照错误类型进行分组，并统计各分组的频数，如表 3 所示。

Table 3. Problem license plate frequency table
表 3. 问题车牌频数表

Plate	N
格式错误	2179
无车牌	1563
未识别	1782
00000000	409

为了充分利用问题车牌数据，基于相邻检测节点的时距关系，通过模糊匹配中的编辑距离算法来校核车牌，默认两车牌字符相似度在 0.8 以上可视为同一个车牌，统计车牌模糊匹配前后问题车牌数如图 4 所示。

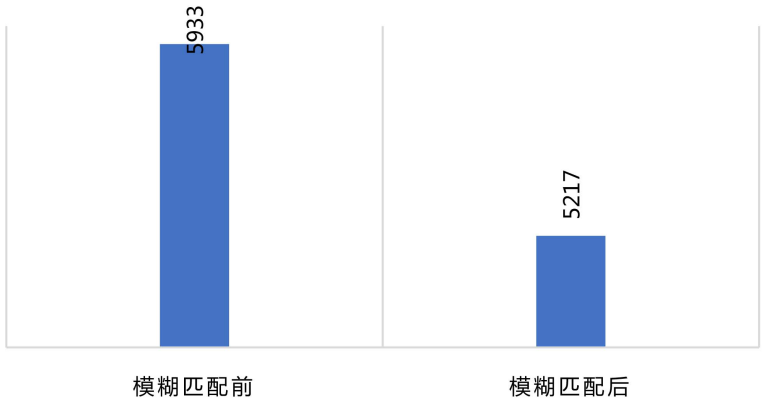

Figure 4. Statistics on the number of license plates before and after the fuzzy matching of license plates
图 4. 车牌模糊匹配前后问题车牌数统计

对于经过模糊匹配后仍存在问题的车牌数据采取删除操作，此外，由于交叉口红灯时间、停车和设备检测数据的重复上传等导致车辆检测数据重复出现，需要剔除重复记录[9]，主要根据：1) 数据中存在完全一样的记录，所有属性均相同；2) 两个相邻的记录相同，除了时间有差异，设置时间差阈值 10。本文使用 sql 语句进行识别：

```
--select
--a.*, DATEDIFF(SECOND, b.time, a.time) as ac, cast(b.direction_id as int)-cast(b.direction_id as int) as bc, cast(b.detection_id as bigint)-cast(a.detection_id as bigint) as cc,
--isnull((nullif(b.licenseplate_id, a.licenseplate_id)), 0) as dc
--into plate--from
--(select ROW_NUMBER() over(order by licenseplate_id, time) as ida, * from plate) as a
--left join
--(select ROW_NUMBER() over(order by licenseplate_id, time)+1 as idb, * from plate) as b
--on a.ida=b.idb
--delete from plate
--where bc=0 AND cc=0 AND dc like '0'
--OR ac<10
```

经过上述处理过程后，即完成车牌照识别数据的预处理工作，需要注意的是，本文主要针对问题车牌照识别数据的识别、校核、清洗方法进行研究，然而，数据的预处理过程需要根据实际的任务需求灵活变化，面对不同问题时，我们应该根据具体问题的特点设计适合的数据预处理方法。

5. 结语

本文通过研究车牌照识别数据的特点和错误情况，设计了一种针对车牌照识别数据的预处理方法，并通过桂林市秀峰区车牌照识别数据证明了该方法的可操作性。该方法基于 mysql 数据库和 C#实现，通过对车牌照识别数据的集成、归约、清理等操作，完成问题车牌照数据的预处理过程。

基金项目

广西高校大学生创新创业项目(No. 201710595051)。

参考文献

[1] 兰昊晖. 车牌识别系统应用场景识别率的研究[J]. 信息通信, 2015(2): 12-13.

[2] 余丰茹, 单飞, 张晓楠, 等. 基于全车牌识别数据的高速公路交通拥挤识别[J]. 公路与汽运, 2014(3): 56-58.

[3] 刘泉. 基于车牌照识别的交通流数据处理平台建立及统计分析[D]: [硕士学位论文]. 北京: 北京交通大学, 2009.

[4] 李瑞敏, 陈熙怡, 张睿博. 基于路口转弯流量的 OD 估计方法研究[J]. 交通运输系统工程与信息, 2015, 15(6): 170-176.

[5] 刘晴辉. VLPR 检纠错支持下的群体出行规律分析与车辆路径追踪研究[D]: [硕士学位论文]. 杭州: 浙江工业大学, 2017.

[6] 阮树斌, 王福建, 马东方, 等. 基于车牌识别数据的机动车出行轨迹提取算法[J]. 浙江大学学报(工学版), 2018, 52, 337(5): 23-31.

[7] 窦志伟. 基于车牌识别数据的交通流参数短时预测[D]: [硕士学位论文]. 成都: 西南交通大学, 2016.

[8] 王龙飞. 基于车牌照的车辆出行轨迹分析方法与实践研究[D]: [博士学位论文]. 西安: 长安大学, 2011.

[9] 杨帅, 于海洋. 基于卡口数据的车辆出行轨迹重构方法研究[C]//第十一届中国智能交通年会大会论文集. 2016.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2326-3431，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：ojtt@hanspub.org