文章编号: 1674-7070(2011)03-0238-06

基于规则引擎的多元大气信息数据质量检查方法

王兴[1] 朱定真[2] 苗春生[1]

摘要

为了提高气象观测资料的质量,为天气预报以及各类气象业务系统提供可靠的数据来源,提出了一种基于规则引擎的气象观测资料质量检查方法. 该方法能对各类常规气象报文进行实时、准确地解码与译码,检查、订正不符合世界气象组织和中国气象局编码规范的报文,并具备灵活的人机交互能力,以适应各类气象观测资料不断发展、变化的要求. 该方法的应用,可取代传统的以气候极值检验为主要手段的质量检查方法,提高了实时资料处理的效率,并显著提升了相关气象产品结果的准确性. 经实践证明,该方法的应用是行之有效的.

关键词

气象观测资料; 质量检查; 规则引擎

中图分类号 P413

文献标志码 A

0 引言

随着气象科学事业发展的广泛深入,人们需要更高质量的气象资料. 然而,由于气象观测资料受到观测环境、硬件技术限制以及人为错误等诸多因素的影响,使得资料的准确性大打折扣. 当前,国内常用的气象观测资料质量检查方法主要包括: 报文格式检查、气候极值检查、内部相关性检查、结合气象学公式的时、空相关性检查以及统计学检查等. 这些方法的应用在很大程度上提高了气象观测资料的准确性[1].

由于这些方法相互独立,一些方法之间存在重复检查或漏查,从而降低了质量检查的效率. 特别是随着气象观测资料种类的丰富以及观测频度的增加,现今每天的数据量已多达几 GB 甚至几十 GB,因此,对实时资料处理的效率提出了更高的要求; 同时,由于这些质量检查方法通常是在计算机程序中写“死”,程序一旦交付实际应用,就很难随意地修改或添加检查的规则和参数. 例如,电子观测设备因长期使用造成的连续性数值漂移,目前只能通过人工调校或手工订正等方法解决[2].

针对此现状,本文提出一种基于规则引擎的气象观测资料质量检查方法.

规则引擎起源于基于规则的专家系统,属于人工智能的范畴,它模仿人类的推理方式,使用试探性的方法进行推理,并使用人类能够理解的术语解释和证明它的推理结论. 规则引擎的应用简化了人类向计算机表述复杂业务逻辑的过程,它通过规则文件来存储业务逻辑,又通过对规则文件的解析来处理业务,从而实现业务逻辑与处理逻辑的分离. 任何一种规则引擎都需要算法的支持,其中,RETE 算法是当前效率最高的一种前向链推理算法之一,本文中的规则引擎即选用该算法[3].

应用基于 RETE 算法的规则引擎对气象观测资料进行质量检查,不仅能够高效地检查、过滤报文中不规范的数据,还能够充分结合气候极值检查、内部相关性检查,以及气象学公式检查等一系列质量检查方法,并提供简捷的质量检查规则编写接口.

收稿日期 2010-10-08

资助项目 国家科技支撑计划(2007BAC29B-0604)

作者简介

王兴,男,硕士,研究方向为气象信息安全技术. sweetdreamworks @ nuist. edu. cn

1 南京信息工程大学 大气科学学院,南京,210044

2 北京华风气象影视信息集团公司,北京,100081

1 数据预处理

目前,绝大多数常规气象观测资料的基本形式是以 ANSI 编码的气象报告. 在实际业务工作中,这些报告并不能直接应用,而必须先经过相应的解报程序对其解码,还原成实际的观测值. 与传统报文解码方式不同,本文提出的基于规则引擎的质量检查方法,首先要对报文进行数据预处理,主要流程包括:报文的归整与报文的拆分两个步骤.

1.1 报文的归整

报文归整的过程是将以"="分隔各个测站数据的报文,重新以行为单位进行分隔,并过滤掉多余的分隔符和无效字符. 归整方法如表 1 所示.

表 1 报文的归整

Table 1 Categorization of meteorological message

原始报文
BBXX
DBLK 05151 99788 10064 41092 /3507 10025 20025 4? 274 55001 7 ////22262 04060 =
TFEA NIL =
LDWR 05151 99660 10018 41497 70511 10108 20065 40161 58003 70222 875//22200 04125 10705 =
归整后的报文
BBXX DBLK 05151 99788...(略去)...22262 04060 =
BBXX LDWR 05151 99660...(略去)...04125 10705 =

表 1 所示,是对报文中一小段海洋地面观测报文进行的归整[4-5]. 经归整后的报文,相较原始报文具有如下优点.

1) 过滤无效的报文. 如原报中"TFEA NIL ="缺少有效的观测数据,故直接滤去.

2) 便于后继程序的处理. 每个测站的数据独占一行,便于程序对报文的解析和拆分.

3) 提高处理效率. 直接将表示报文类别的标识,如"BBXX"添加到各测站(各行)的最前端,从而避免后继程序为查找报文类别而不断地往返文件指针,因此,可以处理效率.

1.2 报文的拆分

根据 WMO-NO. 306《Manual on lodes》的相关规范,通常每个测站的报文每 5 个字符为 1 组(也有例外,如 BBXX 中的船舶称号),组与组之间用一个空格符分隔. 每组数据都具有特定的含义,包含了 1~3 个气象要素信息,而每组数据的首字符或前 2 位字符,通常又表示该组数据所表示的气象要素类型. 报文拆分即是将每行(每个测站)的记录,以分隔符(包括任意多个空格、Tab 及其他特殊字符)为单位进行拆分,并将拆分后的数据填充到指定的单元格中[6]. 在物理实现上是一张内存表,该内存表的结构如表 2 所示. BX00—BX14 表示对应于海洋地面观测报 BBXX 的内存表中各个表项的代号,这些代号是进行质量检查规则编写的重要标识. 第 2 行和第 6 行是各个表项的含义说明;第 3 行和第 7 行是表 1 归整后的报文中,第 1 行报文的拆分结果;第 4 行和第 8 行是表 1 归整后的报文中,第 2 行报文的拆分结果. 其中,第 7 行 BX05 列,本应填充"7////",但由于该数值没有什么实际意义,因此填充时将该组数据丢弃不填. 同理,对于一些明显无意义或前后矛盾的数据项,也采取直接丢弃不填.

表 2 是针对 BBXX 设计的内存表,其他诸如 AAXX、PPAA、TTBB 等报文,虽然各组数据的含义各有不同,但拆分的思路相似. 经拆分后的报文,每组数据都有特定的标识,如 BX00,这些标识将直接应用到动态规则库的规则定义中.

表 2 用以存储拆分数据的内存表

Table 2 Memory table for saving resolution data

BX00	BX01	BX02	BX03	BX04	BX05	BX06	BX07
CallSign	$YYGGi_w$	$99L_aL_aL_a$	$Q_cL_oL_oL_oL_o$	$4i_xhVV$	Nddff	00fff	$1S_nTTT$
DBLK	05151	99788	10064	410920	/3507		10025
LDWR	05151	99660	10018	41497	70511		1010
BX08	BX09	BX10	BX11	BX12	BX13	BX14	…
$2S_nT_dT_dT_d$	$3P_0P_0P_0P_0$	4PPPP	5appp	$6RRRt_R$	$7wwW_1W_2$	$8N_hC_LC_MC_H$	…
20025		4? 274	55001				…
20065		40161	58003		70222		…

2 基于规则引擎的质量检查方法

在逻辑功能上，规则库用于记录各项质量检查的规则. 在物理结构上，规则库是一张或多张二维数据表，这些表可存储于文件或数据库中. 规则库的构建主要包括：规则库的定义和规则的编写两个方面[7].

2.1 规则库的定义

规则库记录了各种对报文进行质量检查的规则，少则几十条，多则几百甚至上千条. 为了提高规则库的可维护性，提升相关程序的处理性能，将这些规则以数据库表的形式进行存储. 无论选用哪种数据库系统，其数据库表的逻辑结构都如图 1 所示.

由图 1 可知，不同的检查方式决定了程序调用相应的处理模块，并且，根据所采用的检查方式，可选择的逻辑关系类型也随之改变. 检查表达式 1 用于设定经拆分后的数据项，或由数据项参与的运算表达式. 检查表达式 2 除具有表达式 1 的编写方式外，还可以编写特定规范的正则表达式. 本文中的正则表达式使用 C#语言，如拆分的正则表达式为“@"\s+"”. 异常处理方式用于指定当满足相应规则后，处理程序所采取的策略[8].

2.2 规则的编写

根据图 1 的表结构，编写气象观测资料的质量检查规则，如表 3 所示.

2.2.1 基于规则库的报文有效性检查

报文有效性检查遵照 WMO-NO.306《Manual on code》中相关编码规范. 以海洋地面观测报为例，该类报文应遵循 FM13-XI Ext 编码规范.

1) 记录完整性检查. BBXX 是海洋地面观测报告的识别字码，YYGG 分别表示测报的日期和正点时间. 由于 YYGG 表明了观测的实际时间，如果报文中缺少该项，则视该条记录无效. 故编写规则如表 3 中 0101 号规则. 与此类似的，还有 $99L_aL_aL_a$、$Q_cL_oL_oL_oL_o$、$4i_xhVV$ 和 Nddff 等数据项也需要进行完整性检查.

2) 非法字符检查. $1S_nTTT$、$2S_nT_dT_dT_d$、$3P_0P_0P_0P_0$、4PPPP、5appp、$6RRRt_R$、$7wwW_1W_2$、$8N_hC_LC_MC_H$ 分别记录了测站的温度. 露点温度、本站气压、海平面气压、过去 3 h 气压变化、过去 6 h 降水量，以及天气现象、云量等信息. 这些信息都是用阿拉伯数字表示的，

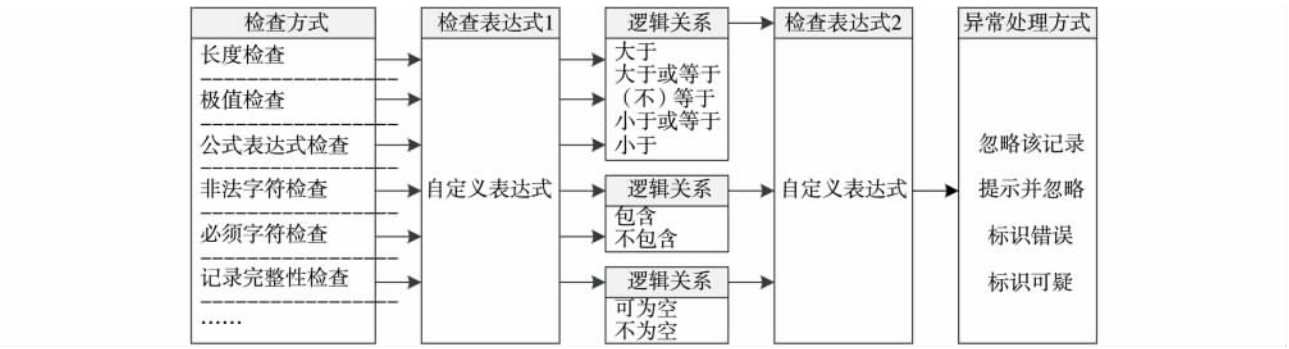

图 1 规则数据库表的逻辑结构

Fig. 1 Logical structure of regular database table

表 3 规则数据库表的示例

Table 3 Demonstration of regular database table

序号	检查方式	检查表达式 1	逻辑关系	检查表达式 2	异常处理方式
0101	记录完整性检查	[BX01]	不为空	(空)	忽略
0102	非法字符检查	[BX10]	包含	@"\D"	提示并忽略
0103	长度检查	[BX07]	不等于	5	提示并忽略
0104	必须字符检查	[BX02]	不包含	@"^99[0-9]{3}$"	提示并忽略
0105	非法字符检查	[BX01]	包含	@"[A-Za-a]"	提示并忽略
0106	最大值检查	getTemp([BX07])	大于	[TableName].[Temp_MaxVaule]×0.8+0.5	标识可疑
0107	最小值检查	getTemp([BX07])	小于	[TableName].[Temp_MinValue]×0.8+0.5	标识可疑
0108	内部一致性检查	getTemp([BX07])	小于	getTemp([BX08])	标识错误

而在表2第7行BX10单元,出现了"4? 274"一项数据.4PPPP表示的内容本应该是测站的平均海平面气压,若为"40274",表示气压值为1 027.4 hPa,而对于"4? 274",若不作检查,处理程序将抛出异常.因此,对于这类错误,可编写如表3中0102号规则."@"\D""是C#中的正则表达式,其含义是:检查指定字符串([BX10])中是否包含非阿拉伯数字的字符.该正则表达式还可编写为如下形式:@"^[0-9]{5}$"或@"^\d{5}$".其中,"^"表示从第1个字符开始进行匹配,"$"表示一直匹配到最后一个字符[9].

3)缺损的数据项检查.通常每一项数据都是由5个字符组成,而在表2第4行BX07单元,出现了"1010"一项数据.对这样的数据项,如果不作检查,程序将无法正确地识别该项数据表示的含义.因此,对于这类错误,可编写如表3中0103号规则.该项规则的含义是:如果[BX07]项的数据,其长度不等于5,处理程序将忽略该条记录,不作处理,并在交互界面或程序运行日志中给出提示.

4)冗余的数据项检查.通过编写规则不仅能够检查缺损的数据项,还能够检查出冗余的数据项.如表2第3行BX04单元,"410920"一项数据,包含6个字符,也可通过长度检查进行识别和过滤.

尽管各类报文的编码方式都不相同,但规则的编写方法基本一致.除上述列举的4种检查外,基于规则库的报文有效性检查还能够实现对报文的时间一致性、报文体积有效性,以及重复报文的检查等[10].

2.2.2 基于规则库的气候极值检查

气候极值检查是众多气象资料质量检查方法中应用最广泛的一种.该方法从纯数值的角度,将各个气象要素的数值与历史资料库中的同期最大值、最小值和平均值进行比较,如果数值在一定的极值区间内,则初步认定该数值正确,反之,则认为数据有错或可疑,进而作进一步的检查.当前,气候极值检查方法主要以独立的程序或作为某一气象业务系统中的模块加以应用,但是,随着观测年份的不断增多,观测数据量的不断增大,以及各种极端天气现象的频发,极值的区间也在不断地扩大.固有的不可调节的程序或模块,已不能适应复杂多变的气候环境.

运用基于规则库的气候极值检查,可以简单、便捷地配置气象要素与历史资料库中数值比较的算法.规则编写的示例如表3中0106号规则.检查表达式1中,getTemp()是一个预定义的函数,其功能是将[BX07]单元所表示的数据,如"10025"译码为直观的温度2.5.检查表达式2中,[TableName].[Column]分别用于指定历史资料库的极值表名称和表中列的名称(或列的顺列号),而历史资料库的物理位置和访问方式需在程序相关配置中进行设置,在规则表达式2中,只需指定表名和列名,其后的"×0.8+0.5",是根据用户特定需求自定义的表达式.其含义是当[BX07]单元所表示的温度值超出历史同期温度极大、极小区间的80%时,认定当前值可疑.用算术表达式表示为"$T_{min}\times 0.8+0.5\leqslant T\leqslant T_{max}\times 0.8+0.5$".考虑到近年来全球气温逐年升高的影响,将比较区间调高0.5 ℃[11].该条规则在规则表中的"逻辑关系"一项没有实质意义,可任意填写或缺省为空.

通过该方法,还可以对诸如本站气压、海平面气压、降水量等各个气象要素进行极值检查,并可根据实际需求自定义极值比较的表达式.

2.2.3 基于规则库的内部相关性检查

内部相关性检查主要用于检查同一条测站记录中,各气象要素自身变化是否合理,不同要素之间是否符合某种物理联系[12].如露点温度≤温度,总云量≥低云量等.其中,表示露点温度≤温度的规则,如表3中0107所示.

2.2.4 基于规则库的其他质量检查方法

由于气象观测资料在预处理阶段被拆分到一组单元中,因此只要运用单元格名称(如上文中的"BX01")和程序预先设定的多种检查方式,即可编写出灵活多变,适应各种需求的气象资料质量检验规则,如基于气象学公式的时间变化规律一致性检查、空间变化规律一致性检查,以及结合统计相关理论而进行的更加复杂精细的检查等.

2.3 规则库与规则引擎

规则引擎采用产生式规则作为基本的知识表达方式.产生式规则的一般形式如:

```
RULE "< Rule Name >"
    < Attribute > < Value >
    < When >
        < LHS >
    < Then >
        < RHS >
END
```

每一条规则都是由一个或多个属性及对应的值、一个Left Hand Side (LHS)和一个Right Hand

Side (RHS) 组成. 每条规则的 LHS 由条件元素和列组成,可以用一阶逻辑或命题逻辑进行表述. 列通常用来表示对一个事实的域约束. LHS 可以由一个或多个条件组成. 当且仅当所有的 LHS 都满足并为真时,RHS 才被执行.

目前规则引擎采用的算法主要有 RETE、LEAPS、LINEAR 和 TREAT 等. 由于 RETE 算法是当前效率最高的一种前向链推理算法之一,而气象业务的数据量大,且实时性要求高[13],因此本文选用该算法. 其核心思想是将分离的匹配项,根据内容动态地构造匹配树,以达到降低计算量的效果. 由于篇幅有限,本文略过 RETE 相关概念,直接举例说明基于 RETE 算法的规则引擎处理如表 3 所列规则的网络结构,如图 2 所示.

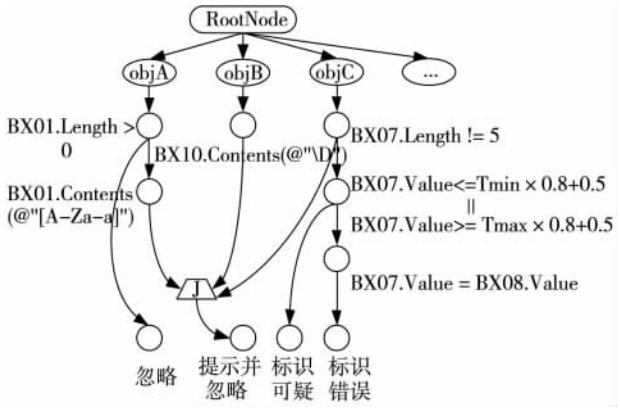

图 2 RETE 网络结构

Fig. 2 A Network Architecture Map of RETE

当一个规则断言一条报文记录(对象),引擎将数据传递到根节点,从根节点进入对象类型节点并沿着网络向下传递. 当数据匹配某个节点的条件,则该节点就将对象记录到相应的 Working Memory 中. 虽然记忆全部匹配的对象或部分匹配的对象需要存储空间,但可以很大程度上提高效率,并具有较好可伸缩性. 各条质量检查规则均按照如图 2 所示的推理过程进行匹配,直到所有报文记录都完成该匹配过程,则规则引擎执行完毕.

通过上述 RETE 算法及网络结构图,运用 Drools 规则引擎工具编写 XML 或 drl 格式的规则代码,以 jar 包的形式发布,也可以作为功能模块添加到其他程序或系统中. 最后再进行规则文件的编译、部署和运行即可完成基于规则引擎的气象资料质量控制过程.

3 结论

基于规则引擎的气象观测资料质量检查方法,能够高效地检查、过滤报文中不符合 WMO 相关规范的数据. 该方法不仅能够充分结合气候极值检查、内部相关性检查,以及气象学公式检查等一系列质量检查方法,还提供简单、便捷的质量检查规则编写接口,实现规则库动态交互能力,以适应气象报文编码规范及质量检查方法不断发展变化的要求.

应用该理论方法开发的常规气象资料实时处理系统,已在南京信息工程大学气象台投入实际应用,运行近一年时间,为天气预报以及 Micaps 等气象业务系统提供可靠的数据来源,并减少了手工操作的复杂性. 经实践证明,该方法的应用是行之有效的.

参考文献

References

[1] 黄刚,屈侠,王鹏飞. 气象数据分析和诊断可视化平台的设计和构想及其在互联网络上的实现[J]. 大气科学学报,2010,33(2):153-159
HUANG Gang,QU Xia,WANG Pengfei. Design of meteorological data analysis and diagnosis visualization system and its realization on the internet [J]. Transactions of Atmospheric Sciences,2010,33(2):153-159

[2] 高华云,应显勋,高峰,等. 气象观测报告的解码规则与算法[M]. 北京: 气象出版社,2006:5-10
GAO Huayun,YING Xianxun,GAO Feng,et al. Decoding rules and algorithm of meteorological observations report [M]. Beijing: China Meteorological Press,2006:5-10

[3] 成敏. 基于规则引擎的动态工作流模型研究与设计[D]. 武汉: 武汉理工大学计算机科学与技术学院,2009
CHENG Min. Research and design of the dynamic workflow model based on the rule engine [D]. Wuhan: College of Computer Science & Technology,Wuhan University of Technology,2009

[4] 中国气象局监测网络司. 地面气象电码手册[M]. 北京: 气象出版社,2007:3-23
China Meteorological Administration. Code manual of surface meteorological [M]. Beijing: China Meteorological Press,2007:3-23

[5] 香港天文台. 船舶天气报告电码[EB/OL]. (2003-05-04)[2010-08-30]. http://www.weather.gov.hk/wservice/tsheet/pms/shipcode_c.htm
Hong Kong Observatory. Code for ship weather reports [EB/OL]. (2003-05-04)[2010-08-30]. http://www.weather.gov.hk/wservice/tsheet/pms/shipcode_c.htm

[6] WMO. WMO publication No. 306: Manual on codes [EB/OL]. [2010-09-16]. http://www.wmo.int/pages/prog/www/WMOCodes/ManualCodes.html

[7] Lott N,Baldwin R,Jones P. The FCC integrated surface hourly database: A new resource of global climate data [M]. National Climatic Data Center Technical Report No. 2001 (01). Asheville: National Climatic Data Cen-

ter,2001
[8] Microsoft MSDN Library. Visual C# 2008 develop [EB/OL]. [2010-07-18]. http://msdn2.microsoft.com/zh-cn/library/
[9] 郭凯红,李文立. 基于规则的大规模试卷文本语块识别方法的研究 [J]. 计算机应用研究,2009,26(4):1391-1393
GUO Kaihong, LI Wenli. Study of massive paper texts chunking based on rules [J]. Application Research of Computers,2009,26(4):1391-1393
[10] 任芝花,刘小宁,杨文霞. 极端异常气象资料的综合性质量控制与分析 [J]. 气象学报,2005,63(4):526-533
REN Zhihua, LIU Xiaoning, YANG Wenxia. Complex quality control and analysis of extremely abnormal meteorological data [J]. Acta Meteorologica Sinica,2005,63(4):526-533
[11] National Climatic Data Center. Global historical climatology network (GHCN) quality control of monthly temperature data [EB/OL]. [2010-08-24]. http://www.ncdc.noaa.gov
[12] 任慧龙,谷富生,张雪梅. 浅析地面气象数据的质量控制 [J]. 山西气象,2004(3):45-46
REN Huilong, GU Fusheng, ZHANG Xuemei. Brief analysis of quality control of ground meteorological data [J]. Shanxi Meteorological Quarterly,2004(3):45-46
[13] 刘金龙. Drools 规则引擎模式匹配效率优化研究及实现 [D]. 成都:西南交通大学信息科学与技术学院,2007
LIU Jinlong. Research and implementation on efficiency optimization of pattern-matching in drools rule engine [D]. Chengdu: School of Information Science & Technology, Southwest Jiao Tong University,2007

Research on quality check method of polytropic atmospheric data based on rule engine

WANG Xing[1] ZHU Dingzhen[2] MIAO Chunsheng[1]
1 School of Atmospheric Sciences, Nanjing University of Information Science & Technology, Nanjing 210044
2 Huafeng Group of Meteorological Audio & Video Information, Beijing 100081

Abstract In order to improve the quality of meteorological data and provide reliable data sources for weather forecasts and other meteorological applications, the paper presents a quality check method of meteorological observational data based on rule engine. The method is able to decode and translate accurately the routine meteorological message in real time. It can check and revise the message which does not meet the norms published by the WMO and the CMA. It also has flexible ability in man-machine interaction, by which to satisfy the developing and changing requirements about meteorological observational data. This method may replace the traditional quality control method, which use the meteorological extreme value examination. It can enhance the processing efficiency of real-time data, and can also promote the accuracy of result produced by meteorology applications. The practical applications prove that this method is effective and efficient.

Key words meteorological observational data; quality check; rule engine