文章编号：1006-4354（2011）05-0035-03

基于气象科学数据库的数据提取及统计应用

燕东渭，杨　艳

（1. 陕西省气象信息中心，西安　710014；2. 陕西省大气探测技术保障中心，西安　710014）

摘　要：陕西省气象科学数据共享服务系统是中国气象科学数据共享服务网的一个重要组成部分。针对公共气象服务业务对历史气象数据的需求，以陕西气象科学数据共享服务系统数据库为基础，给出了对数据库按类别实施批量统计均值、极值和排序等三类查询的通用方法。方法可以大大提高用户交互式查询数据库的效率，在气象行业外其他数据库同样有广泛的实用价值。

关键词：气象科学数据库；均值；极值；排序

中图分类号：P409　　**文献标识码**：B

气象科学数据共享服务系统是中国气象局牵头的国家科技部基础性建设重点项目，是覆盖全国30个省的分布式数据服务系统[1-3]，2006年已基本建成。目前多数省级共享系统提供仅限于Web方式的原始数据下载，基于该数据库的其他应用和服务还未建立起来。在“陕西省公共气象服务平台”软件系统开发过程中，以陕西省气象科学数据共享服务系统数据库为基础，总结出在给定年限内，对诸要素均值、极值统计和排序的高效方法。

1　方法

气象工作者常需参考较长时间序列的统计数据来开展公共气象服务业务。气象行业通用标准时间跨度是30 a，通常需将30 a的统计数据制作成数据集使用。对陕西省气象科学共享数据库原始历史数据二次加工，批量提取各气象站30 a气象要素的平均值、极值以及历史排名等统计数据，将统计结果按统一格式写入“公共气象服务平台”数据库，供数据查询时使用。

数据源采用的是Microsoft SQL Server数据库系统，库中包含陕西省各地面气象站建站以来日、旬、月、年的气象要素数据表格。气温、降水等要素的均值和极值的统计，可依靠运行SQL语句实现，大体可分为提取、处理和存储三步实现。首先，用Microsoft SQL Server客户端软件中的查询分析器，逐一将需统计的要素值、站

收稿日期：2010-11-18

作者简介：燕东渭（1975—），男，陕西周至人，硕士，气象电子高工，从事气象信息技术和网络管理等。

基金项目：气象科学数据共享中心—省级数据资源建设与共享服务（2005DKA31700-06-13）

（3）雷达频繁报仰角－终限位故障，仰角数码管显示仰角处于－2°以下。该现象说明雷达天线处于下限位状态，将伺服机柜上的抱闸信号置于断开状态，然后人工将天线仰角推于0°以上，再将抱闸信号置于闭合状态，该故障排除。此故障主要是因为雷达的抱闸信号出现故障，导致天线掉下。抱闸信号通过汇流环传输，若雷达频繁在某个固定的方位角掉下，基本可判定为汇流环信号接触不好所致，清洗汇流环即可。

参考文献：

[1]　潘新民．新一代天气雷达（CINRAD/SB）技术特点和维护、维修方法［M］．北京：气象出版社，2009.

[2]　何炳文，顾松山，高嵩，等．伦茨伺服控制器的功能及其在CINRAD/SB中的应用［J］．气象，2006，32（7）：52-57.

[3]　北京敏视达雷达有限公司．中国新一代多普勒天气雷达CINRAD/CB用户手册［G］．北京：北京敏视达公司，2006.

号及时间等信息提取出来，按照统一顺序临时存放在 Excel 表格。因为总站数和站号顺序都是固定的，所以每次提取数据的总条数和顺序也是一致的。之后根据不同要素数据的特点，将可合并的要素合并到一张 Excel 表格中。最后用 Microsoft SQL Server 客户端软件中的数据导入导出工具，将表格导入“公共气象服务平台”数据库。平均数据和极值数据需根据其特点，分别采用不同的 SQL 语句提取。SQL 查询的设计是解决问题的关键。

2 数据的提取和统计

2.1 统计均值

SQL 语言统计均值、极值和求和等一般采用聚合函数，和其它函数的最大区别在于它们一般作用在多条记录上。聚合函数和 GROUP BY 子句配合，可让聚合函数对属于不同组的数据分别起作用，高效制作各类报表，还方便实现一次性统计不同站、不同时间均值。

提取全省各站 30 a 各月平均气温和降水等信息 SQL 语句（SURF _ CLI _ SN _ MUL _ MON 为月数据表名称，V12001 _ 501 要素的 32766 是特殊值，表示该要素缺测，统计均值时需剔除）

```
SELECT V01000, V04002, AVG (V12001 _ 501) AS V1, AVG (V13011) AS V2
FROM dbo. SURF _ CLI _ SN _ MUL _ MON AS a
WHERE (V04001 <2001) AND (V04001 >1970) AND (V12001 _ 501<>32766)
GROUP BY V01000, V04002 ORDER BY V01000, V04002
```

类似，可统计全省各站不同时间某气象要素超过给定阈值的次数，如各站历史上各月降水超过 100 mm 的次数统计

```
SELECT V01000, V04002, COUNT (*) AS number FROM dbo. SURF _ CLI _ SN _ MUL _ MON
WHERE (V04001<2001) AND (V04001 >1970) AND (V13011>1000)
GROUP BY V01000, V04002 ORDER BY V01000, V04002
```

2.2 统计极值

极值统计相对复杂一些，不仅要提取极值本身，而且要考虑极值出现的时间及并列极值等问题。若用聚合函数 Max () 虽能直接求极大值，却无法准确定位极值在数据表格中的位置，无法找到极值对应的时间等信息。通常可利用子查询先查出极值出现的年份，同时从表中检索该年份及相应的极值。该方法通常可行，但若出现并列极值的情况，子查询会因返回的年份值不止 1 个而报错。使用 SQL 语句给表格和自身建立内联接[4]，可筛选出并列的极值记录。

为了方便数据应用，针对极值多次出现的情况，不但应将极值重复的时间都提取出来，而且最好在生成的报表中合并成一行。因此，极值统计须分两步完成。第一步和提取均值类似，先用查询分析器，逐一将站号、要素极值及时间等信息用特定的 SQL 语句提取出来，按照统一的顺序存放在临时表中。对于月数据的极值来说，先站号排序，再时间排序，最后是要素值和出现的年份，一条信息一行。不同要素极值无法一次性提取，须逐一提取。

用内联接实现 30 a 月极端最高气温的提取

```
SELECT a. V01000, a. V04002, a. V04001, T. V
FROM dbo. SURF _ CLI _ SN _ MUL _ MON AS a INNER JOIN
( SELECT V01000, V04002, MAX (V12211 _ 505) AS V
FROM dbo. SURF _ CLI _ SN _ MUL _ MON WHERE (V04001<2001) AND (V04001 >1970) GROUP BY V01000, V04002) AS T
ON a. V01000 = T. V01000 AND a. V04002 = T. V04002 AND a. V12211 _ 505 = T. V
WHERE (a. V04001 < 2001) AND (a. V04001 > 1970) ORDER BY a. V01000, a. V04002
```

第二步是对第一步查询结果的再处理。该结果中并列极值的记录是分别出现，为了保证存储时和其他数据结构的一致性，需将并列极值的数

据合并成一行，即并列极值出现的年份合并成一个字段，不同年份间用点号隔开。为了直接用SQL语句将查询结果中并列极值的多行合并成一行，可使用SELECT语句的FOR XML子句并指定XML模式，对数据库执行SQL查询，可以XML格式返回结果。检索XML格式的结果，可像处理字符串一样对其做一些特殊处理，完成并列极值出现年份的合并。在第一步得到的包含并列极值的临时表（表格名是＃TMP）基础上，引入XML，可合并月数据的极值，合并后的并列极值年份间用点号分隔。

SELECT DISTINCT V01000，V04002，V04001＝

STUFF（（SELECT'.'＋V04001 FROM ＃TMP WHERE V01000＝t.V01000 AND V04002 ＝ t.V04002 AND V12211 _ 505 ＝ t.V12211 _ 505 FOR XML PATH（''））,1,1,''），V12211 _ 505 FROM ＃TMP t

处理极值时，虽然将要素极值出现的年份合并，但无法将不同要素极值同年份的合并一并解决，因此不同要素的极值须保存在不同的表格中，即使其长度和顺序一致。

2.3　排序

各站不同要素的历史前十位排名数据的提取比极值提取更加复杂，提取的不只是极值，还有排名。实质上是一个复杂的分组排序问题，处理方式与极值提取有较大差异。首先须解决分组排序问题，其次是序号提取。SQL Server 2005引入了row _ number、rank、dense _ rank等排序函数[4]，可实现多种排序。这几个函数都会根据order子句指定字段的顺序为查询出来的每行记录生成一个序号，只是序号的生成稍有区别。dense _ rank函数可为排序字段值相同的行分配相同的序号，同时保证序号的连续性。因此用dense _ rank函数最适合。dense _ rank（）语法DENSE _ RANK（）OVER（[partition by list1] order by list2）

参数partition by将FROM子句所生成的结果集划分为多个将要应用dense _ rank函数的分区，允许为查询出的行集的每个分区（组）分别计算排名[4]，类似group by子句功能，满足排序问题按站号和时间的不同分别排名的要求。参数order by确定将dense _ rank值应用于分区中各行的顺序。

需要注意的是，通常提取的日最高、最低气温和最大降水等要素的数据的总量是不一致的，若某站降水数据为0，进行历史排名无意义（且此时并列排名可能较多），这样的信息没有必要提取；还有闰月问题，自1951年有气象数据记录以来，很多站2月29日的数据不到10个。所以不同要素需独立保存在不同表格中。若只提取各站各要素的第一名，相当于极值提取。

提取单站日最高气温前十位数据（V01000，V04001，V04002，V04003，V12211分别表示站号、年份、月份、日期和日平均气温，SURF _ CLI _ SN _ MUL _ DAY是日数据表）。

SELECT ＊ FROM（SELECT S.V01000，dense _ rank（）OVER（partition BY S.V01000，S.V04002，S.V04003 ORDER BY S.V12211 DESC）AS rank，S.V04002，S.V04003，S.V12211，S.V04001 FROM SURF _ CLI _ SN _ MUL _ DAY AS S）AS T WHERE T.rank＜11 ORDER BY T.V01000，T.V04002，T.V04003，T.rank

3　结语

随着公共气象服务业务的不断发展，各种需求越来越多，要实现更多个性化的信息提取服务，单靠执行SQL语句可能无法完全满足需求，还需要开发专门的软件系统。

参考文献：

[1]　杨青军．省级气象科学数据共享系统设计与实现[J]．中国西部科技，2008，7（7）：27-28.

[2]　李永花．青海省气象科学数据共享服务系统的设计与实现[J]．青海气象，2008，（S1）：117-119.

[3]　王国复，李集明，邓莉，等．中国气象科学数据共享服务网总体设计与建设[J]．应用气象学报，2004，15（增刊）：21-26.

[4]　Itzik Ben-gan，Dejan Sarka，Roger Wolter. Microsoft SQL Server 2005技术内幕：T-SQL查询[M]．赵立东，译．北京：电子工业出版社，2007.