DOI:10.13878/j.cnki.jnuist.2019.01.019

成勤[1,2]　肖稳安[2]　王清龙[3]　项建国[1]　于乃莲[1]　陈华[1]

正则表达式在闪电定位资料处理中的应用

摘要

随着闪电定位观测手段的提高以及观测仪器的改进,闪电定位资料的质量也在不断提高,其应用也越来越广泛.原始闪电定位资料从数据整理、入库、查询、分析都离不开正则表达式这种功能强大的形式语言.将正则表达式应用于闪电定位资料处理中,可以大大简化处理过程,提高工作效率.本文通过闪电定位资料入库实例,介绍了在.NET 平台下,基于 C#的正则表达式在闪电定位资料处理中的应用.

关键词

闪电定位资料;正则表达式;.NET 平台;C#语言

中图分类号 P427.32+1

文献标志码 A

收稿日期 2017-10-22

资助项目 湖北省气象局科技发展基金(2014Z01)

作者简介

成勤,女,硕士,主要研究方向为雷电监测预警及相关软件开发.chengqin3@163.com

1 宜昌市防雷中心,宜昌,443000

2 南京信息工程大学 大气物理学院,南京,210044

3 宜昌市气象局,宜昌,443000

0 引言

近年来,随着雷电远程定位理论的成熟和雷电监测网的建设[1],各类雷电属性记录更加完备、雷电定位更加精确、雷电记录更加全面,闪电定位数据也越来越受到国内外学者的重视.闪电定位资料被广泛用于研究闪电活动规律[2]和分布特征[3]以及与其他气象要素的关系,可为改进防雷措施提供依据[4].

常见的原始闪电定位资料是.TXT 格式的文件,文件名为当天的日期,文件中记录闪电发生的时间、纬度、经度、强度、陡度、定位方式、误差等信息.闪电定位系统每天产生一个文件,随着时间的增长,闪电定位数据越来越多,闪电参数(例如雷电流平均强度、陡度、闪电密度等信息)的查询也越来越繁琐.

正则表达式的强大功能逐渐被人们认识,并在很多领域有广泛的应用[5-8].基于正则表达式的高性能算法[9-11]为研究人员解决了数据处理上的难题.闪电定位资料作为一种监测数据,也可以利用正则表达式来处理和分析.

1 正则表达式基础

正则表达式[12](Regular Expression、regex 或 regexp,缩写为 RE),也译为正规表示法、常规表示法,在计算机科学中,是指一个用来描述或者匹配一系列符合某个句法规则的字符串的单个字符串,就是用一个“字符串”来描述一个特征,然后去验证另一个“字符串”是否符合这个特征.由普通字符(如字符 a 到 z)以及特殊字符(称为元字符)组成的文字模式作为一个模板,将某个字符模式与所搜索的字符串进行匹配.正则表达式可以:

1)验证字符串是否符合指定特征,比如验证是否是合法的邮件地址;

2)查找字符串,从一个长的文本中查找符合指定特征的字符串,比查找固定字符串更加灵活方便;

3)替换,比普通的替换更强大.

正则表达式这个概念最初是由 Unix 中的工具软件(例如 sed 和 grep)普及开的,经过几十年的发展,正则表达式的强大功能逐渐被人们认识,特别是在字符串处理方面.正则表达式逐渐从模糊而深奥的数学概念,发展成为在计算机各类工具和软件包应用中的主要功能.

不仅众多 Unix 工具,其他主流操作系统(例如 Linux、Windws、HP、BeOS 等)也支持正则表达式,目前主流的开发语言(PHP、C#、Java、C++、VB、Javascript、Ruby 等)以及数以亿万计的各种应用软件中,都可以看到正则表达式的应用实例[13].

2 闪电定位资料的正则表达式

应用正则表达式处理数据的关键在于写出符合数据特性的正则表达式.也许对于初学者来说,正则表达式是杂乱无序的字符,难写难读,但是其功能强大,应用广泛,比常用的 String.Substring( )方法更准确有效.当资料的各字段之间的分隔符各异,或各条记录的格式不统一时,利用一般字符串截取的方法很难准确、完整地提取信息,难以实现数据的批量处理,而利用正则表达式不仅可以准确匹配各个字段,还能提高工作效率.

如图 1 所示,同一列数据的长度不一定相同,列与列之间的分隔符各异.以提取“定位方式”相关信息为例,普通字符串截取的方法很难准确、完整地提取信息,而正则表达式可以实现这一功能,所匹配的正则表达式为:(? <=定位方式:).*?.

2.1 .NET 平台下的正则表达式

正则表达式提供了功能强大、灵活、高效的方法来处理文本.正则表达式的全面模式匹配表示法可以快速分析大量文本以找到特定的字符模式,提取、编辑、替换或删除文本子字符串,或将提取的字符串添加到集合以生成报告.对于处理字符串(例如 HTML 处理、日志文件分析和 HTTP 标头分析)的许多应用程序而言,正则表达式是不可缺少的工具.Microsoft.NET Framework 正则表达式并入了其他正则表达式实现的最常见功能(例如在 Perl 和 awk 中提供的功能),不仅与 Perl 5 正则表达式兼容,还包括一些在其他实现中尚未提供的功能,例如从右到左匹配和即时编译..NET Framework 正则表达式类是基类库的一部分,它们可以和面向公共语言运行库的任何语言或工具(包括 ASP.NET 和 Visual Studio 2010)一起使用.C#的正则表达式以 Perl 与 regexp 为基础,包括惰性定量符(??, *?, +?, {n, m}?)、正向和逆向前瞻(look ahead)以及条件估值.

基类库命名空间 System.Text.RegularExpressions 是所有与正则表达式相关的.NET 框架对象的大本营.支持正则表达式的核心类是 Regex[14],Regex 构造函数有 4 个重载函数:

1)Regex ( ),用于初始化 Regex 类的新实例;

2)Regex (String),为指定的正则表达式初始化并编译 Regex 类的一个新实例;

3)Regex (SerializationInfo,StreamingContext),使用序列化数据初始化 Regex 类的新实例;

4)Regex (String,RegexOptions),用修改模式选项为指定的正则表达式初始化,并编译 Regex 类的一个新实例.

在使用 C#语言编写闪电定位资料的正则表达式之前,需要在开发项目中添加正则表达式的命名空间,即在代码里加入“using System.Text.RegularExpressions;”.

2.2 闪电定位资料的正则表达式

根据闪电定位资料的使用性质可知,常用的闪电数据包括闪电发生的时间、经纬度、雷电流强度与陡度.闪电定位资料中通常还包括行号、闪电定位方式、误差、能量、电荷量等信息.分析闪电定位数据的属性,编写各闪电参数信息的正则表达式如下:

1)匹配日期和时间的正则表达式 rgx1:

```
Regex rgx1=new Regex(@"(? <=^\d+)\d{4}-\d{2}-\d{2} \d{2}:\d{2}:\d{2}\.? \d*",RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出字符串中形如“yyyy-mm-ddhh:ii:ss.aa”的字符串(其中 y、m、d、h、i、s、a 都是数字),表示闪电记录中日期和时间,例如:“2009-12-09 23:45:12.234”.

①匹配年份的正则表达式 rgxNian 为

```
Regex rgxNian=new Regex(@"\d{4}(? =-\d{2}-\d{2})",RegexOptions.IgnorePatternWhitespace | RegexOptions.Multiline);
```

它的含义是找出闪电记录中形如“yyyy-mm-dd”字符

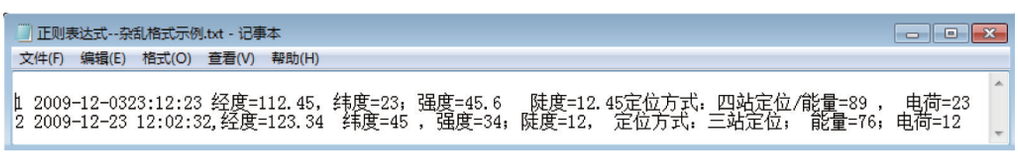

图 1 闪电定位资料杂乱数据格式示例

Fig. 1 Example of lightning location data in different formats

串中的“yyyy”字符串,“yyyy”为数字,表示闪电记录中的年份.

②匹配月份的正则表达式 rgxYue 为

```
Regex rgxYue=new Regex(@"(?<=\d{4}-)\d{2}(?=-\d{2})",RegexOptions.IgnorePatternWhitespace | RegexOptions.Multiline);
```

它的含义是找出闪电记录中形如“yyyy-mm-dd”字符串中的“mm”字符串,“mm”为数字,表示闪电记录中的月份.

③匹配日期的正则表达式 rgxRi 为

```
Regex rgxRi=new Regex(@"(?<=\d{4}-\d{2}-)\d{2}", RegexOptions.IgnorePatternWhitespace | RegexOptions.Multiline);
```

它的含义是找出闪电记录中形如“yyyy-mm-dd”字符串中的“dd”,“dd”为数字,表示闪电记录中的日期.

2)匹配纬度的正则表达式 rgx2:

```
Regex rgx2=new Regex(@"(?<=纬度=)[\d|-]\d+\.?\d*", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出闪电记录中表示纬度的字符串,例如“纬度=32.12”.

3)匹配经度的正则表达式 rgx3:

```
Regex rgx3=new Regex(@"(?<=经度=)[\d|-]\d+\.?\d*", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出闪电记录中表示经度的字符串,例如“经度=112.34”.

4)匹配雷电流强度的正则表达式 rgx4:

```
Regex rgx4=new Regex(@"(?<=强度=)[\d|-]\d+\.?\d*", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出闪电记录中表示雷电流强度的字符串,例如“强度=15”.

5)匹配雷电流陡度的正则表达式 rgx5:

```
Regex rgx5=new Regex(@"(?<=陡度=)[\d|-]\d+\.?\d*", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出闪电记录中表示雷电流陡度的字符串,例如“陡度=10”.

6)其他闪电信息的正则表达式:

不同的闪电定位系统产生的闪电定位数据的字段数和格式可能有所差别,有的闪电定位资料中还可能包括以下闪电信息:闪电记录的行号、定位方式、定位误差、电荷量、能量等.

①匹配行号的正则表达式 rgx6 为

```
Regex rgx6 = new Regex(@"^\d*", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是在字符串的开始处找出一个或多个数字,表示一条闪电记录的行号.

②匹配闪电定位方式的正则表达式 rgx7 为

```
Regex rgx7 = new Regex(@"(?<=定位方式:).*?", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出字符串中以“定位方式:”开头,后接一个或多个特殊字符(包括汉字)的一段字符串,表示闪电记录中的定位方式,例如“定位方式:三站定位”.

③匹配误差的正则表达式 rgx8 为

```
Regex rgx8=new Regex(@"(?<=误差=)[\d|-]\d+\.?\d*", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出字符串中以“误差=”开头,后接一个小数(小数的位数为两个或以上)的一段字符串,表示闪电记录中的定位误差,例如“误差=23.12”.

④匹配能量的正则表达式 rgx9 为

```
Regex rgx9=new Regex(@"(?<=能量=)[\d|-]\d+\.+\d+", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出字符串中以“能量=”开头,后接一个小数(小数的位数为两个或以上)的一段字符串,表示闪电记录中的能量,例如“能量=800”.

⑤匹配电荷量的正则表达式 rgx10 为

```
Regex rgx10=new Regex(@"(?<=电荷=)[\d|-]\d+\.+\d+", RegexOptions.Compiled | RegexOptions.Multiline);
```

它的含义是找出字符串中以“电荷=”开头,后接一个小数(小数的位数为 2 个或以上)的一段字符串,表示闪电记录中的电荷量,例如“电荷=100”.

以上代码均指定将正则表达式编译为程序集,加快执行速度,并开启多行模式,分别在任意一行的行首和行尾匹配,而不仅仅在整个字符串的开头和结尾匹配.

3 正则表达式应用实例

正则表达式是一种功能强大的形式语言,贯穿于闪电定位资料处理的始末,包括资料整理、入库、查询、分析、结果显示等环节.以下将以数据入库为例,介绍正则表达式在闪电定位资料处理中的应用.

将闪电定位数据入库,不仅能大大简化查询数据的步骤,提高分析和处理资料的速度,而且还有如下优点:

1)实现数据共享;

2)减少数据冗余,维护数据的一致性;

3)保证数据的独立性;

4)数据实现集中控制;

5)数据一致性和可维护性,以确保数据的安全性和可靠性;

6)故障恢复.

由数据库管理系统提供一套方法,可及时发现故障和修复故障,从而防止数据被破坏.数据库系统能尽快恢复运行时出现的故障,可能是物理上或是逻辑上的错误,比如对系统的误操作造成的数据错误等.

3.1 数据库设计

根据数据的使用性质,选取需要的参数导入 Microsoft SQL Server 2008 数据库中,主要包括闪电发生的时间(year,month,day)、经纬度(lon,lat)以及雷电流强度与陡度(qiangdu,doudu),其数据属性如表1所示.

表1 闪电参数数据属性

Table 1 Data attributes of lightning parameters

列名	数据类型	允许 Null 值
year	int	□
month	int	□
day	int	□
lon	float	□
lat	float	□
qiangdu	float	□
doudu	float	□

3.2 数据入库程序

以下为利用正则表达式实现闪电定位资料入库的主要程序.程序中的“src”表示数据存放的路径,“sqlYearName”表示数据的年份.首先利用正则表达式提取闪电信息,再将提取的信息插入到数据表中,规定数据类型.具体代码如下:

```
static public bool onelinktodb(string src,string sqlYearName)
    {    using (sqlconn)
        {  sqlconn.Open();
           using (SqlCommand sqlc = sqlconn.CreateCommand())
           {  string st,sql;
              //部分代码省略
              if (File.Exists(src))
              {  FileStream fs;
                 StreamReader sr;
                 Regex rgxNian = new Regex(@"\d{4}(?=-\d{2}-\d{2})",RegexOptions.IgnorePatternWhitespace | RegexOptions.Multiline);
                 Regex rgxYue = new Regex(@"(?<=\d{4}-)\d{2}(?=-\d{2})",RegexOptions.IgnorePatternWhitespace | RegexOptions.Multiline);
                 Regex rgxRi = new Regex(@"(?<=\d{4}-\d{2}-)\d{2}",RegexOptions.IgnorePatternWhitespace | RegexOptions.Multiline);
                 Regex rgx2 = new Regex(@"(?<=纬度=)[\d|-]\d+\.+\d+",RegexOptions.Compiled | RegexOptions.Multiline);
                 Regex rgx3 = new Regex(@"(?<=经度=)[\d|-]\d+\.+\d+",RegexOptions.Compiled | RegexOptions.Multiline);
                 Regex rgx4 = new Regex(@"(?<=强度=)[\d|-]\d+\.+\d+",RegexOptions.Compiled | RegexOptions.Multiline);
                 Regex rgx5 = new Regex(@"(?<=陡度=)[\d|-]\d+\.+\d+",RegexOptions.Compiled | RegexOptions.Multiline);
                 fs = new FileStream(src,FileMode.Open,FileAccess.Read);
                 sr = new StreamReader(fs);
                 while ((st = sr.ReadLine()) != null)
                 {  //部分代码省略
                    sql = @"INSERT INTO [lightning].[dbo].["+sqlYearName+@"]
                          ([year]
                          ,[month]
                          ,[day]
                          ,[lat]
                          ,[lon]
                          ,[qiangdu]
                          ,[doudu]
                          )
                    VALUES
    ("+makeint(rgxNian.Match(st).ToString())+@"
    ,"+makeint(rgxYue.Match(st).ToString())+@"
    ,"+makeint(rgxRi.Match(st).ToString())+@"
    ,"+makefloat(rgx2.Match(st).ToString())+@"
    ,"+makefloat(rgx3.Match(st).ToString())+@"
    ,"+makefloat(rgx4.Match(st).ToString())+@"
    ,"+makefloat(rgx5.Match(st).ToString())+@")";
    sqlc.CommandText = sql;
    sqlc.ExecuteNonQuery();//部分代码省略}}}}}
```

通过以上分析可知,正则表达式构造出相应的

匹配模式,从用户自然语言的内容中找出对应的关键字,并把其转化成数据库可以识别的标准格式,这样有利于提高数据的录入效率.

4 结束语

不同地区的闪电定位资料数据格式可能稍有差别,但是一般都会包含闪电发生时间、地点、雷电流强度、陡度等信息,且其结构基本相同.将正则表达式引人到测量数据处理中是非常有意义的.利用正则表达式匹配闪电定位资料的各条记录,分析和处理闪电定位资料,不仅可以提到工作效率,还可以提高准确率.由于正则表达式非常灵活,对于同一种功能需求,可以有多种不同的写法,在实践中可以结合自己的习惯来编写.闪电定位资料的正则表达式写法不是唯一的.本文给出的一种闪电定位资料的正则表达式,可供资料处理者参考.

正则表达式的语法晦涩难懂,表面上看起来杂乱无章,很难掌握,常令许多程序员望而生畏,敬而远之.然而,一旦熟练掌握了正则表达式之后,便能获得很大的益处,不仅能在进行字符数字处理时节约大量编程的时间,极大地提高工作的效率,同时也能使所编写出来的程序代码得到很好的优化.相信正则表达式在今后海量的测量数据处理中的应用会越来越开阔.

参考文献

References

[1] 陈家宏,张勤,冯万兴,等.中国电网雷电定位系统与雷电监测网[J].高电压技术,2008,34(3):425-431
CHEN Jiahong, ZHANG Qin, FENG Wanxing, et al. Lightning location system and lightning detection network of China power grid[J].High Voltage Engineering,2008,34(3):425-431

[2] Orville R E, Huffines G R. Cloud-to-ground lightning in the united states: NLDN results in the first decade,1989-1998[J]. Monthly Weather Review, 2001, 129(5): 1117-1193

[3] 李永福,司马文霞,陈林,等.基于雷电定位数据的雷电流参数随海拔变化规律[J].高电压技术,2011,37(7): 1634-1641
LI Yongfu,SIMA Wenxia,CHEN Lin,et al.Law between parameters of lightning current and elevation based on lightning detection data[J]. High Voltage Engineering, 2011,37(7): 1634-1641

[4] Zoro R, Suhana H. Improvement of lightning protection system on distribution lines: a case study at South-Jakarta,Indonesia[J].Proc of Seminar National Ketenagalistrikan,2004,SN-112

[5] Yu F, Chen Z F, Diao Y L, et al. Fast and memory efficient regular expression matching for deep packet inspection[C]// Proc of the 2006 ACM/IEEE Symposium on Architectures for Networking and Communications Systems,2006: 93-102

[6] 杨波,张立娜.基于C#正则表达式的农业文献管理系统的研究与应用[J].安徽农业科学,2012,40(5): 2988-2990
YANG Bo,ZHANG Lina.Research and application of C# regular expression in agricultural document management system[J].Journal of Auhui Agricultural Science,2012, 40(5): 2988-2990

[7] 谭玉玲.基于正则表达式的数据处理应用[J].武汉理工大学学报(信息与管理工程版),2010,32(2): 249-252
TAN Yuling. Applications of regular-expression based data processing[J].Journal of WUT(Information & Management Engineering),2010,32(2): 249-252

[8] 许光,黄宏志,刘娜.正则表达式在Web数据验证中的优化机制研究[J].计算机与数字工程,2011,39(4): 50-52,157
XU Guang,HUANG Hongzhi,LIU Na.Research on optimization mechanism of regular expression in web data validation[J].Computer & Digital Engineering,2011,39(4): 50-52,157

[9] 金军航,张大方,黄昆.高性能正则表达式匹配算法评估[J].计算机工程,2010,36(19): 269-271
JIN Junhang, ZHANG Dafang, HUANG Kun. Evaluation of high-performance regular expression matching algorithms[J]. Computer Engineering, 2010, 36(19): 269-271

[10] 邓绪斌.基于最优树联配的正则表达式学习算法[J].复旦学报(自然科学版),2011,50(6): 797-802
DENG Xubin. Learning regular expressions via optimal tree alignment[J]. Journal of Fudan University(Natural Science),2011,50(6): 797-802

[11] 徐克付,齐德昱,郑伟平,等.一种基于Bloom Filter的正则表达式集合快速搜索算法[J].华南理工大学学报(自然科学版),2009,37(4):37-41
XU Kefu,QI Deyu,ZHENG Weiping,et al.A fast regular expression set matching algorithm based on bloom filter[J]. Journal of South China University of Technology(Natural Science Edition),2009,37(4):37-41

[12] Watt Andrew.正则表达式入门经典[M].北京:清华大学出版社,2008
Watt Andrew.Beginning regular expression[M].Beijing: Tsinghua University Press,2008

[13] Goyvaerts Jan, Levithan Steven.正则表达式经典实例[M].北京:人民邮电出版社,2010
Goyvaerts Jan,Levithan Steven.Regular expression cookbook[M].Beijing: Posts & Telecom Press,2010

[14] Liger Francois,McQueen Craig,Wilton Paul.C #字符串和正则表达式参考手册[M].北京:清华大学出版社,2003

Liger Francois,McQueen Craig,Wilton Paul.C # text manipulation strings handling and regular expressions handbook[M].Beijing: Tsinghua University Press,2003

Application of regular expressions in data processing of lightning detection and location system

CHENG Qin[1,2] XIAO Wenan[2] WANG Qinglong[3] XIANG Jianguo[1] YU Nailian[1] CHEN Hua[1]

1 Yichang Lightning Protection Center of Hubei Province,Yichang 443000

2 School of Atmospheric Physics,Nanjing University of Information Science & Technology,Nanjing 210044

3 Yichang Weather Bureau of Hubei Province,Yichang 443000

Abstract With the continuous improvement in lightning observation method and instrument,lightning detection and location data are substantially improved in quality and have found wide applications.As a powerful formal language,regular expression has been extensively applied to process original lightning detection and location data,from data compilation and storage to data analysis and query.The regular expression can simplify the processing procedures and improve working efficiencies in data processing of lightning detection and location system.This paper introduces the application of C# based regular expression in lightning detection and location data processing on the. Net platform.

Key words data of lightning detection and location system;regular expression;.NET;C#