统计、去除文件中的重复行Linux认证考试 PDF转换可能丢失图片或格式，建议阅读原文 https://www.100test.com/kao_ti2020/607/2021_2022__E7_BB_9F_E8_AE_A1_E3_80_81_E5_c103_607065.htm 在数据库中，如果表中记录有重复的话，则只需要通过distinct关键字就可以达到去除重复行的目的。那么在Unix操作系统中，是否也有便利的工具能够实现这个需求呢?答案当然是肯定的。为了实现这个需求，我们需要用到sort排序命令和uniq去除重复行命令。 众所周知，在Unix系统维护中，系统工程师经常需要把多个文件合并成一个文件。此时就会遇到一个问题，即将多个文件连接或者合并在一起的时候，可能会产生重复的记录。这是系统工程师不希望看到的。那么该如何消除这些重复的记录呢?熟悉排序命令的读者，一定知道利用sort排序命中一个-u 可选项可以达到删除重复行的目的。但是这个功能并不是很强。如不能够帮助系统工程师找出哪些行是重复的或者统计重复行出现的次数等等。要实现这些复杂的功能，就需要借助于这个uniq去除重复行命令了。 第一步：先对文件中的记录进行排序。 如现在系统工程师将两个文件合并后，产生如下一个文件log.txt。 销售部出现系统备份故障 采购部出现系统备份故障 销售部出现系统备份故障 财务部出现系统备份故障 这是一个系统备份程序出现故障时的提示信息。为了便于管理，系统工程师往往先将一个星期或者半个月的错误记录合并在同一个文件中。然后再把重复的记录去除掉，就可以发现哪些部分的备份出了问题。如果不去掉重复行的话，那么一个个核对过去就会很麻烦。看着一长串错误列表(其实很多都是重复的)，也不知道该如何下手。另外统计重复行

出现的次数(每个部分系统备份故障程序的次数)也可以帮助系统工程师判断这个错误是不是偶然性的。为此，现在系统工程师主要想实现三个功能。第一是能够统计这个文件中错误信息重复的次数.第二是能够知道哪些故障信息出现了一次以上.第三就是得到一个去除了重复行的文件，以方便工程师分析问题原因并最终解决问题。这三个功能的话利用uniq命令都可以轻松解决。不过需要注意的是，在使用这个去除重复行的命令之前，必须要先对这个记录文件进行排序。其实这个原理跟数据库中的distinct去除重复行的关键字工作原理是类似的。在数据库中采用这个关键字去除重复记录时，数据库会自动对相关的记录进行排序，然后再去除重复行。而现在这个uniq命令自身没有排序的功能。这也是Unix操作系统的一个特点，即每个命令只完成单一的功能。而这个去除重复行的uniq命令对于那些没有经过排序的命令是不起任何作用的。如上面这个日志文件，如果文件中的记录没有排序，那么这个uniq命令就无法去除这个重复的记录或者统计重复记录出现的次数。为此系统工程师要做的第一步，就是对这个文件排序。如可以使用sort log.txtgt.。但是这个uniq命令则不同。因为这个命令本省就可以带两个文件名作为参数。如uniq 原文件目标文件。注意这个格式并不是说让操作系统分别对原文件与目标文件都进行去除重复行的动作。而是对原文件进行去除重复行的操作，然后将执行的结果保存到目标文件中。也就是说，此时uniq命令只是对原文件进行处理，并不会对目标文件进行任何去除重复行的操作。换一句话说，uniq命令一次只能够对一个文件进行去除重复行的操作。另外就是需要注意的是，将执行的结果保存到一个文件中

不需要用到重定向符号。 第三步：将以上两个步骤合二为一。 如果按照上面那个步骤来进行操作，虽然是可行的，思路也比较清晰。但是中间会多一个排序生成的过渡文件。等操作完成后，需要手工将这个文件删除。显然这增加了工作量。用过管道符的读者一定知道，这个管道符有一个很特殊的用途。即将某个命令的执行结果传递给下一个命令，让其作为下一个命令的参数。如可以使用命令sort log.lst | uniq loguniq.lst。 这个命令是什么意思呢?首先是利用sort命令对log.lst文件中的记录进行排序。然后将排序的结果传递给uniq命令。最好操作系统会将去除重复行后的记录保存到loguniq.lst文件中。注意，这里将执行的结果保存到文件中也没有使用这个&gt.重定向符号。而是采用了“ -文件名 ”这个可选项。这个对于那些熟悉重定向符号的读者可能看起来不怎么舒服。但是这是Unix系统中文件中的一个例外，各位系统工程师只需要记住即可。利用上面这个命令，来代替上面两个步骤，除了工作量减轻不少以外，最重要的就是不会产生中间的垃圾文件。即这个命令执行的过程中，不会产生排序后的文件。 第四步：选取重复的行或者统计重复行出现的次数。 以上三个步骤只是完成了一项功能，即去除文件中重复的记录。如果需要对文件中重复的记录进行统计，或者只显示重复的行，则各位读者还需要关注一下笔者下面给大家讲述的内容。要实现这两个功能，就需要用到三个可选项，分别为-u、-d与-c。其中-u表示只让操作系统显示没有重复的行。-d是告诉操作系统只显示重复的记录。而可选项-c是让系统统计重复行出现的次数，并且会在记录的前面加一列内容，表示重复行出现的次数。注意，系统工程师也可以将

这些结果保存到一个文件中。但是同上面的原理一样，不需要使用管道符号来实现这个目的。可以直接在原文件后面加入一个目标文件，来实现保存结果的目的。 第五步：截取特定的列来去除重复记录。 如上面这个文件中，不止一个列。如上面这个文件，其内容如下： 销售部出现系统备份故障 故障日期2009年6月3日 采购部出现系统备份故障 故障日期2009年6月5日 销售部出现系统备份故障 故障日期2009年6月6日 此时文件中的记录是没有重复的。但是系统工程师想获得的信息是哪几个部门在重复的出现类似的问题。此时系统工程师需要先截取某一个列，如第一列中的信息。然后再将第一列中的内容利用uniq命令来去除重复的行或者统计重复行出现的次数。要实现这个目的的话，也比较简单。只是在上面几个步骤中，多了一个前期的准备工作而已。即通过cut等类似的命令，将文件中的某列内容截取出来，保存到一个文件中。然后再进行排序与去除重复行的操作。如果系统工程师不想产生中间文件，也可以利用管道符将cut、sort、uniq等三个命令连接起来。另外如果以后需要多次用到这个功能，工程师想贪图方便的话，还可以将这个命令以别名的形式保存起来。如此的话，下次直接将这个命令的组合当作系统命令来使用即可。