

2023 年

“认证杯”数学中国数学建模网络挑战赛

第二阶段

B 题 考订文本

古代文本在传抄过程中,往往会出现种种错误,以至于一部书可能流传下来多种版本。在文献学中,错误往往被总结成“讹”、“脱”、“衍”、“倒”等形式,也可能同时出现多种错误。错误可以在传抄过程中不断累加。

1. “讹”是指对原始文本的篡改。包括无意中写错单个文字,也包括根据传抄者自己的理解篡改完整的词汇,句子乃至整段内容。例如《红楼梦》中著名菜肴“茄鲞”的做法,就有不同版本的古籍流传至今,而且内容相去甚远,其中势必存在被传抄者篡改的部分;
2. “脱”是指误删文字。包括遗漏单个文字或者成段内容。例如《荀子·劝学》一文中“蓬生麻中,不扶而直”一句,在古籍的传世版本中并无后文。后经清代王念孙考证,后面应有“白沙在涅,与之俱黑”一句;
3. “衍”是指误增文字。包括误增单字或词,误增整句的情况也有。例如三国人物“士仁”在《三国演义》通行本中写作“傅士仁”,有人猜测这个“傅”姓本是衍字而来。增整段的情况较少,往往是传抄者将其他文献或自己原创的批注加进文本,后世无法辨识所致;
4. “倒”一般是指交换原有文字的位置。单个文字位置对换往往是由于传抄失误,大段乃至整篇文字的对换往往是由于装订失误。例如明代于谦诗作《石灰吟》中有“粉骨碎身浑不怕”一句,在一些传抄版本中被误作“粉身碎骨浑不怕”。

不仅是古代的传抄者会出错,即使是现代的通信或存储设备,当一条信息被多次转发或转录以后,也无法避免随机发生的错误。在此,我们将此问题改造成更加理想化的形式:假设原始文本的长度足够大,而且在传抄过程中,传抄者并不和其他版本进行互相校核。这样,在足够长的流传或转发过程中,不同的错误叠加,就可能会产生大量不同的版本。请你建立合理的数学模型,研究如下问题。

第一阶段问题:

1. 请你设计合理的方案,衡量两个不同版本的文本之间的差异大小。
2. 如果一个版本是从另一个版本经过多次传抄而来,我们希望估计两个文本之间经历的传抄次数。请分析并解决这个问题。在建模时请注意:为了进行有效的估计,我们还需要知道哪些必需的信息?
3. 在解决前面提出的问题时,有一些方案虽然在概念上很合理,但会遇到实际计算上的困难。现请你针对前两问,分别设计一个有效而快速的算法。请描述算法的原理,估计其速度,并举算例。

第二阶段问题:

1. 如果我们有多个不知年代的抄本,请通过对文本的研究,挖掘它们之间的关系,即每个版本究竟是通过哪个版本传抄而来。请你和你的团队建立合理的数学模型来解决这个问题,并自行构造算例来进行验证。
2. 如果我们虽然有多个后世的抄本,但原本已经失传,请根据对这些抄本的研究,恢复原本最可能的样子。请你和你的团队建立合理的数学模型来解决这个问题,并自行构造算例对方法的效果进行评价。