准确性、全面性、美观性
——测试数据设计中的三要素

杭州外国语学校　杨帆

【关键字】　测试数据　准确性　全面性　美观性

【摘　要】　测试数据是当今信息学竞赛不可或缺的有机组成部分，其作用已经越来越被人们所重视。本文提出了测试数据设计中信、达、雅，即准确性，全面性，美观性三要素的观点，并就此进行了逐一论述。

一、引论

国际信息学奥林匹克竞赛(IOI)自 1989 年创办以来已经举办了整整十届，而全国信息学奥林匹克竞赛(NOI)也已有了 15 年的历史。在这些年的发展过程中，信息学竞赛不断进行着自我完善，探索出了一套有自身特点的方法。信息学竞赛和其他学科竞赛有着很大的不同，其中之一，就是评分的方法不同：一般的学科竞赛，采用的是分步给分的办法，即每一步中间过程均有相应的分数；而信息学竞赛，从早年的 IOI 开始，就实行了黑箱测试法，即不对源程序进行阅读、分析，而仅仅根据源程序对于所给定的测试数据得出的结果的正确性进行评分，所有这一切均由电脑自动完成。与相比原先的白箱测试，黑箱测试显得更为客观、公正、高效，因此目前被普遍采用。在黑箱测试过程中，给定的测试数据起着至关重要的作用，其重要程度，甚至不亚于题目本身。一道题目，即使再优秀，如果没有好的测试数据，其价值将大打折扣；同样，一道看似平凡，或是毫不起眼的题，如果配上卓越的测试数据，往往就能够起到令人拍案叫绝的效果。同时，测试数据也是评判题目难易的一个重要关卡，一道题目，采用不同的测试数据其难易程度会有很大的差别。由此可见，测试数据是信息学竞赛题中不可或缺的有机组成部分。因此，对测试数据的讨论是重要的，也是必要的。

著名学者严复曾经在《天演论·译例言》中提出“译事三难信达雅”，把信、达、雅作为了翻译的标准。尽管文学翻译和信息学竞赛并没有必然联系，但是严复的这个三字标准同样可以运用在测试数据的设计上。信，是真实、确切的意思；达，是透彻、通达的意思，可引申为完备；雅，是高尚、美丽的意思。显然，信、达、雅分别对应了测试数据设计的三要素：准确性、全面性和美观性。任何优秀的测试数据，必定是这三者的完美结合。下面，本文将对测试数据设计的信、达、雅三要素进行一一论述。

二、本论

（1）信——测试数据的准确性

显然，测试数据的准确性的重要程度是至高无上的，如果失去了准确性，其他方面就无从谈起，对于一道信息学竞赛题来说，如果包含了错误的测试数据，那么它的竞赛价值就等于零。具体地说，测试数据的准确性体现在以下几个方面：一、测试数据的输入输出格式与题目要求符合；二、测试数据在题目限制的范围之内；三、测试数据的输出结果是正确的。这三点，缺一不可。

首先，测试数据的输入输出格式必须与题目要求符合。输入和输出格式的重要性在实行了计算机自动测试之后充分地显示出来，因为格式的错误和算法的错误造成的后果是相同的，计算机在自动测试过程中只判别选手的输出文件和标准输出文件是否有区别，如果有，则判选手程序错误。因此，如果测试数据中的标准输出文件错误，将会造成所有选手正确的输出均被判错。同时，如果输入数据的格式有误，或者没有按题目要求进行排序等，也会使选手的正确程序无法读出正确数据，从而得到错解。在竞赛中出现这样的事故，后果将不堪设想。

其次，测试数据一定要在题目的限制范围内。所谓题目的限制范围，有很多含义，比如题目规定的数据规模限制，数据的上界和下界，数据的类型等等。例如 IOI'98《天外来客》一题，“输入文件大小可达 2M”是对数据规模的限制，“$0<n\leq 20$，$0<a\leq b\leq 12$”是各数据的上限和下限，而“以 2 表示结尾的 0,1 序列”就是数据的类型。在这点上，任何测试数据都不能越雷池一步。在平时练习时，经常出现为了测试某个程序的运算速度而采用大规模数据的情况，在解搜索题中这种现象更为普遍。这样的测试方法就给测试数据的正确性埋下了很大的隐患。因为这样的数据规模很有可能会超出题目限制，或者表面上符合题目要求，但在程序运算过程中出现了数据越界的情况——这是最容易发生，也是最难被查出的。例如，IOI'98《多边形》一题，题目规定顶点数值总在[-32768,32767]的范围内，而一般的规模稍大的测试数据往往会出现运算结果在范围内，中间数值却在范围外的情况，显然这样的数据就不是符合要求的数据。

最后，也是最重要的，就是测试数据的标准输出结果，即俗称的“标准答案”必须正确无误。这不仅是对信息学竞赛题的要求，而且是对所有学科竞赛题的共同要求，甚至可以说是对所有题目的要求。要实现这点，对于信息学竞赛来说，必须保证标准程序的准确性，因为测试数据的标准输出结果是由标准程序产生的，标准程序的错误将直接导致测试数据标准输出的错误。在设计测试数据的时候，同样应该对这方面加以足够重视，决不能有麻痹思想。

综上所述，只有注意了以上三个方面，测试数据才能保证其必须的正确性。正确的测试数据才是合格的测试数据。

（2）达——测试数据的全面性

严复说过：“顾信而不达，虽译犹不译也。”翻译如此，测试数据的设计同样如此。如果只考虑准确性，还远远不能称得上是好的测试数据，充其量只是符合最低要求罢了。因此，在做到测试数据的准确性之后，必须考虑它的全面性，只有这样才能区分程序的优劣，达到竞赛的目的。

测试数据的全面性，大致体现在两个方面：一是对特殊情况的考查；二是对算法的效率和程序的时空承受能力的考查。这两点缺一不可。下面分别对全面性的这两个方面进行论述。

对特殊情况的考查是十分重要的。所谓特殊情况，又可以分为两种，一种是题目的边界条件，另一种则是题目没有明文规定禁止出现，而又不合常情的情况。这两种情况都是很容易被忽略的，尤其是后者。

先来讨论对边界情况的考查。每道题都有自己的限制条件，即边界条件。显然，当测试数据的范围超过边界条件时，该测试数据失去了准确性。但是，当测试数据的范围恰好在题目的边界条件上时，就达到了对边界条件考查的目的。这样的测试数据，不但是准确的，而且是优秀的。

一般说来，题目的限制范围有两个，即上限和下限。对边界条件的考查，一般情况下都是对题目要求的下限的考查，因为对上限的考查往往需要规模较大的数据，对算法的效率和时空承受能力有较高的要求，可以归为对算法的效率和时空承受能力的考查一类。所以，对边界情况的考查就是对题目限制范围下限的考查。例如 IOI'98《夜空繁星》一题规定"0≤星座总数目≤500"，这样就可以设计一个全空的星图，作为对星座数目等于 0 的边界情况的考查。同样是 IOI'98，《圆桌骑士》一题规定"0≤骑士数目≤63"，可以依此设计没有骑士的数据，考查选手考虑问题的全面性。

科学界有一句名言：非绝对禁止者，皆不无可能。同样，在信息学竞赛中，只要是题目没有明文规定禁止出现的情况，都有可能、且有必要出现在测试数据中。例如 NOI'97《文件匹配》的第 16 个测试数据中就出现了对同一个文件，既要求进行操作，又要求不进行操作的情况。由于题目中并没有明文规定不允许出现这种情况，而这个数据又恰好对此进行了考查，因此是一个漂亮的测试数据。这些特殊情况，既是选手考虑问题时容易忽视的地方，也同样是设计测试数据时容易忽视的地方。对于这些细节，无论选手还是命题者，都必须加以足够重视。

对算法的效率和程序的时空承受能力的考查同样是十分重要的。

一道题目，不同的选手在解答过程中，会建立不同的数学模型，采用不同的算法，并且，这些算法对于本题来说都是可行的。基于这点，在测试时必须从算法效率上区分选手算法的优劣，而这个任务，就理所当然地落到了测试数据身上。

对算法的效率的考查一般使用的方法是设计大规模的测试数据，比如 IOI'98《多边形》一题，数据规模稍大，采用搜索法的程序，运算时间就会成几何级数增长，而基于动态规划算法的程序，效率就比较稳定。显然，通过这样的测试，就达到了区分算法优劣的目的。

必须注意到，选手即使采用了完全相同的算法，在编程时，也会因为各种原因，用不同的数据结构去实现相同的算法，这就必须对程序的时空承受能力进行考查。

在这里，程序的时间承受能力不仅与算法的效率有关，还与程序的优化以及预处理的程度等各种因素有关。这点，在搜索算法中显得尤为突出。为了对算法效率相同的程序进行优劣区分，就必须设计一些测试数据，使进行优化或预处理的程序能够在规定时限内出解，反之则不能，例如 IOI'98《图形周长》一题，可以设计一个矩形完全包含的数据，这样，对此进行预处理的程序显然就占有很大优势。

而程序的空间承受能力与程序采用的数据结构密切相关。采用好的数据结构，往往能够节省内存开销。显然这些在运算时间上不能得到体现。因此，就必须针对这些情况，设计一些测试数据，区分程序数据结构的优劣。比方说，对于搜索题，可以设法增加搜索层数；对于需要处理大量数据的题可以增大数据容量；对于运算量大的题，可以增大数据规模。这样，假如程序采用的数据结构不是最优的，就有可能出现程序执行出错，如栈溢出、堆溢出、超过储存范围等等，或者干脆造成死机。这样一来，对于不同的程序，空间承受能力立刻得到了区分。只有算法效率高，而且数据结构选用恰当的程序才能通过这些测试数据。这样的数据无疑是有极大作用的。

综上所述，要做到全面测试，死板地规定 5 个或 10 个的测试数据显然不行，测试数据的最佳数量是由题目本身决定的，和题目的难度以及要求的数据规模密切相关。少了，不能做到全面测试；多了，又起不到太大作用。例如 IOI'98《圆桌骑士》一题，题目本身比较简单，最大规模的数据也只有 63 个骑士，因此只需少量数据就可以实现全面测试了；而像《图形周长》，不仅数据规模大，而且情况繁多，使用不同的算法效率也将有较大区别，因此就必须采用较多的测试数据进行测试。

同时，测试数据的难度往往能够决定一道竞赛题的难度，这也是需要十分注意的。一般情况下，可以把测试数据的分值作以下分配：

简单数据（规模较小）	25%
特殊情况	20%
对算法效率考查（规模较大）	20%
对程序时空承受能力考查（规模较大）	35%

必须指出的是，对程序时空承受能力考查的测试数据同样也考查了算法效率，因此上面的四部分的划分并不截然分明。按如上比例划分测试数据，显得比较平均，适合一般竞赛。但是，对于层次较高，或是层次较低的竞赛，就应该重新设计测试数据难度比例。请看以下两种划分比例：

	一	二
简单数据（规模较小）	55%	5%
特殊情况	10%	15%
对算法效率考查（规模较大）	20%	25%
对程序时空承受能力考查（规模较大）	15%	55%

显然，对于同一道题来说，第一种测试数据难度比例降低了题目难度，而第二种则提高了题目难度。这样，测试数据就起到了调节题目难度的作用。这样的例子屡见不鲜。例如 NOI'98 第一题《个人所得税》就极为典型。这道题本身并不算难，算法一目了然。它之所以出现在全国竞赛上，而且得分率仅为 11%，就是测试数据在起调节作用。本题的 10 个标准测试数据有 9 个是对程序时空承受能力的考查，只有使用高精度运算才能得出正确解。在竞赛时，只有极少数选手考虑到了这点，通过了 8 到 9 个测试数据，其余选手无一幸免。不妨在此做一个假设：假如当时比赛时并没有采用这批测试数据，而用规模较小的数据，这样，这道题的得分率肯定在 90%以上。即使在分区联赛中，这样的题也只能算是简单题了。

由此可见，测试数据的难度比例是极其重要的。在做到了全面测试之后，测试点分值比例分配将直接影响题目的难度和得分率。当然，这些都是建立在全面的测试数据上的。没有全面性，测试数据就很难影响题目的难度。因此，归根结底，测试数据的全面性才是真正的要点所在。

（3）雅——测试数据的美观性

仅仅做到了准确性和全面性，能不能算得上好的测试数据？答案是否定的。“子曰辞达而已，又曰言之无文，行之不远。三者乃文章正轨，亦即为译事楷模。故信达而外，求其而雅。”同样，对于测试数据来说，“信达而外，求其而雅”也是不可缺少的。所谓测试数据的雅，就是指测试数据的美观性，它和文学的“雅”是有很大区别的。后者指的是语言文字的润色，能使文章更为生动。显然，测试数据只是简简单单的字符、数字的组合，要做到生动的确是勉为其难了，因此，它的“雅”，它的美观，是有自己独特的含义的。但是历来信息学竞赛的命题者对此都没有引起足够的重视。

常言道：规范是美，和谐是美。测试数据的美正是建立在这两点的基础上的。测试数据不仅仅是给电脑用来进行自动测试、自动评分的，它最终还是要给人看的。

选手或是教练在赛后对题目进行分析、总结时，往往要查看测试数据。此时，一个和谐、规范的测试数据起到的作用就远远超过一个杂乱无章的测试数据。

那么，究竟怎么样测试数据才是规范的，才是和谐的呢？不妨先看一个例子，以 IOI'98《图形周长》为例，给出下面五个测试数据（为了表述方便、直观，测试数据均转换成原始图形）：

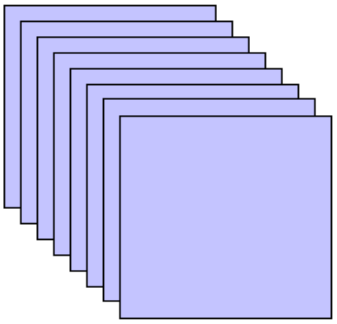

图 1

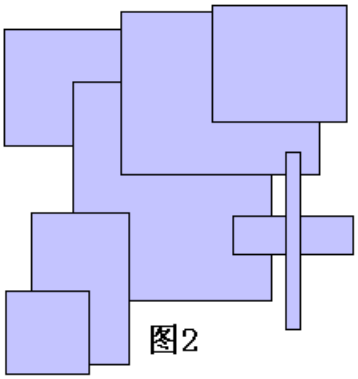

图 2

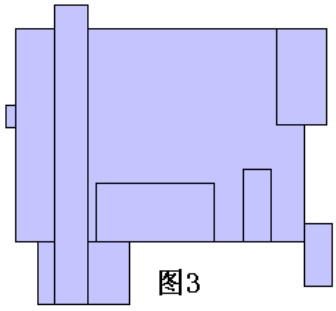

图 3

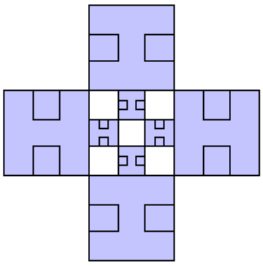

图 4

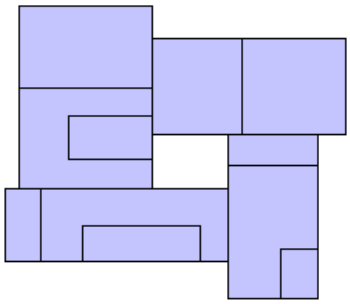

图 5

显然，图 1、图 4 给人的感觉要远远好于图 2、图 3 给人的感觉，而图 5 则介于两者之间。为什么？这就是规范、和谐的作用。图 2、图 3 给人杂乱无章的感觉，已经转换成直观的图形尚且如此，就更不用说原数据了。由此可以看出，规范、和谐都是相对概念，因此没有绝对的规范和绝对的和谐。所谓规范、和谐只是人的一种感觉罢了。如果真要下个定义，可以说假如只看测试数据，就能在脑中形成一个直观的图形，那么这个测试数据就是美的。

对比这个标准，在历年竞赛的测试数据中，能够称得上“雅”的凤毛麟角，更多的数据是随机产生的。不可否认，随机产生的数据具有一般性，在每道题的数据中搀杂几个随机数据的确必要，甚至是达到全面性的必须，但过多过滥的随机数据就不但不是必须，而且破坏了整道题应有的美感。不仅如此，当选手在赛后分析题目的时候，可以想象，如果遇到的都是随机产生的测试数据，他将不可能对程序进行很好的分析，因为他读不懂这些数据；相反，假如他遇到的是美观的数据，那么他就可以更好地，更彻底地分析程序，人脑、电脑一起发动，自然事半功倍。因此，单就这点而言，测试数据的美观性就是极为必要的，因为一道竞赛题的价值不仅仅只在于竞赛，更重要的是在赛后，选手对它分析、总结，从而提高自己的水平。如果没有美观的测试数据，那么题目就不能发挥它在赛后的作用，它的价值也就打了折扣。

由此可见，测试数据的美观性同样显得重要。在今后的竞赛中，对这点的重视程度必将逐年提高。

三、结论

综上所述，信、达、雅是测试数据设计中的三要素，测试数据只有具备了准确性、全面性、美观性，才能称得上是优秀的。但是，这三方面的重要性又互不相同。显然，“信”是基础，“达”是关键，“雅”是提高。没有“信”，测试数据就失去了存在的根本依据，“达”和“雅”就无从谈起，因此，“信”是基础；测试数据光有准确性不够，还必须有全面性，当然全面性不是一个数据就能达到的，需要一组测试数据来实现，只有全面的测试数据才能在竞赛中起到科学测试的作用，因此“达”是关键；一道好的竞赛题，它的价值不仅仅体现在竞赛中，更重要的是能够使选手在竞赛后的分析总结里得到提高，因此测试数据在这个过程中就起到了关键作用，美观的测试数据能最大限度发挥题目潜能，所以“雅”是提高。但是显然，对于一组优秀的测试数据来说，在不顾此失彼的情况下，三者均不可缺少。只有做到了信、达、雅的有机结合，测试数据才是优秀的。

一组优秀的测试数据在题目中起到的作用是不可忽视的，它往往能够对题目的难度产生直接的影响，有时还能够化腐朽为神奇，使一道毫不起眼的题产生惊人的效果。由此可见，测试数据是信息学竞赛题中不可或缺的有机组成部分，必须加以足够重视，进行深入研究。以上只是我个人的一些看法，仅以此抛砖引玉，有关测试数据的设计这个话题，还有待选手和教练们进行更进一步的探讨。