

同样，也不要将字符与表示形式混淆起来。为了让书法作品更好看，很多手写体和字体允许人们把相邻的字符漂亮地连写起来，称为连笔（ligatures），这样两个字符就平滑地连接在一起了。母语为英语的作者常把 F 和 I 结合为 FI 连笔（参见图 16-5a 和图 16-5b），而阿拉伯语的作者常把字符“LAM”和“ALIF”结合为一种很优雅的连笔（参见图 16-5c 和图 16-5d）。

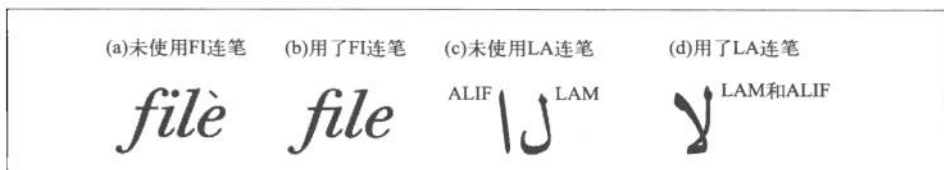


图 16-5 连笔是相邻字符的另一种风格的表示形式，并非新的字符

这里给出一般的规则：如果用一种字形替代另一种的时候，文本的意思变了，那这些字形就是不同的字符。否则，它们就是同一个字符的不同风格的表示形式。⁹

16.3.5 编码后的字符集

根据 RFC 2277 和 2130 的定义，编码后的字符集把整数映射到字符。编码后的字符集经常用数组来实现¹⁰，通过代码数值来索引（参见图 16-6）。数组的元素就是字符¹¹。

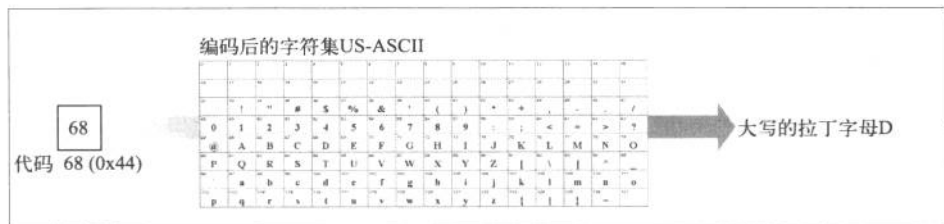


图 16-6 可以把编码后的字符集看作数组，把数值化的代码映射到字符

下面我们来看一些重要的编码后的字符集标准，包括具有历史意义的 US-ASCII 字符集、ASCII 的 iso-8859 扩展、日文的 JIS X 0201 字符集以及统一字符集（Universal Character Set, Unicode）。

注 9：语义和表示方式之间的区别并不总是很清晰的。为了实现的方便，已经为同一个字符的某些表示变体分配了不同的字符。不过我们还是尽量避免这种做法。

注 10：数组可以是多维的，这样代码数字中的不同二进制码就可以索引到数组的不同维。

注 11：图 16-6 使用了一个网格来表示编码后的字符集。网格中的每个元素都包含一个字符图像。这些图像应看作是符号，图像“D”是字符 LATIN CAPITAL LETTER D 的简称，而不是任何特定的图形化字形。