有些字符编码（比如 UTF-8 和 iso-2022-jp）更加复杂，它们是可变长（variable-length）编码，也就是说每个字符的位数都是可变的。这种类型的编码允许使用额外的二进制位表示拥有大量字符的字母表（比如汉语和日语），仅用较少的二进制位来表示标准的拉丁字符。

16.2.2 字符集和编码如何工作

现在我们来看看字符集和编码到底做了什么。

我们想把文档中的二进制码转换为字符以便显示在屏幕上。但由于有很多不同的字母表，也有很多不同的方法把字符编码成二进制码（这些方法各有优缺点），我们需要一种标准方法来描述并应用把二进制码转换为字符的解码算法。

把二进制码转换为字符要经过两个步骤，如图 16-2 所示。

- 在图 16-2a 中，文档中的二进制码被转换成字符代码，它表示了特定编码字符集中某个特定编号的字符。在这个例子里，解码后的字符代码是数字编号 225。
- 在图 16-2b 中，字符代码用于从编码的字符集中选择特定的元素。在 iso-8859-6 中，值 225 对应阿拉伯字母“FEH”。在步骤 a 和 b 中使用的算法取决于 MIME 的 `charset` 标记。

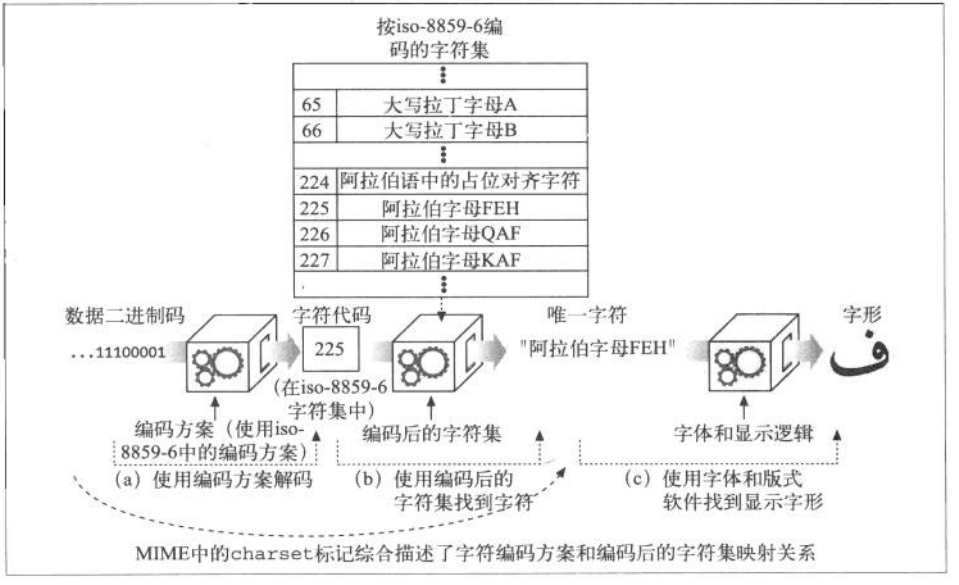

图 16-2 HTTP 协议中的 `charset` 是字符编码方案和编码后字符集的组合