

编码大全

拔赤 bachi@taobao.com

<http://www.uedmagazine.com>

2010-04-23

淘宝网

- 编码之初
- ASCII
- 国际化
- 第二种国际化
- HTTP之URL编码
- HTTP之BASE64编码
- 字体
- 编码之初

- 编码之初
- ASCII
- 国际化
- 第二种国际化
- HTTP之URL编码
- HTTP之BASE64编码
- 字体
- 编码之初

编码之初 – 之初

- 摩尔斯码

- 单位是 “位” （长短音）

- 01 -> A

- 1000 -> B

- 1010 -> C

- 输入法编码

- 单位是 “字符”

- VRM -> 淘（郑码）

- IQRM -> 淘（五笔）

- 4452 -> 淘（区位）=> GB2312

编码之初 – 01编码

- 计算机内码
 - 01编码
 - 编码单位 “字”
 - 2A -> *
 - 30 -> 0
 - 41 -> A
 - 6DD8 -> 淘 (unicode)
 - CCD4 -> 淘 (gbk)
 - B25E -> 淘 (big5)

- 编码之初
- **ASCII**
- 国际化
- 第二种国际化
- HTTP之URL编码
- HTTP之BASE64编码
- 字体
- 编码之初

ASCII – 最初的128个字符

- 长度
 - 一个字（8位）
 - 最高位用作校验位 $2^7 = 128$
- 范围
 - 00 ~ 7F

- 编码之初
- ASCII
- **国际化**
- 第二种国际化
- HTTP之URL编码
- HTTP之BASE64编码
- 字体
- 编码之初

国际化 – 扩充ASCII

- ISO8859 编码标准集合

ASCII

ISO8859-1

德法(Latin-1)

ASCII

ISO8859-2

东欧(Latin-2)

ASCII

ISO8859-6

阿拉伯(Arabic)

ASCII

ISO8859-7

希腊(Greek)

淘宝网

...

ISO8859 (拉丁字符集)

- 子集的不兼容
 - D9 -> Ù (ISO8859-1) 西欧
 - D9 -> Ω (ISO8859-7) 希腊
- 单字节 0 ~ 255
- ISO8859不兼容东亚字符

国际化 – 再次扩充ASCII

- 东亚字符集

ASCII

GB2312-80

简体中文

ASCII

BIG5

繁体中文

ASCII

SJIS

日文

...

东亚字符集 -ISO8859的悲剧重演

- 子集的不兼容
 - B8A1 -> 腹 (BIG5) 繁体中文
 - B8A1 -> 浮 (GB2312) 简体中文
- 双字节
- 与ISO8859各编码集不兼容

导致的悲剧

- 一段文本无法同时使用多种语言
- 一段文本无法同时使用简体和繁体

GBK – 简体/繁体编码的兼容

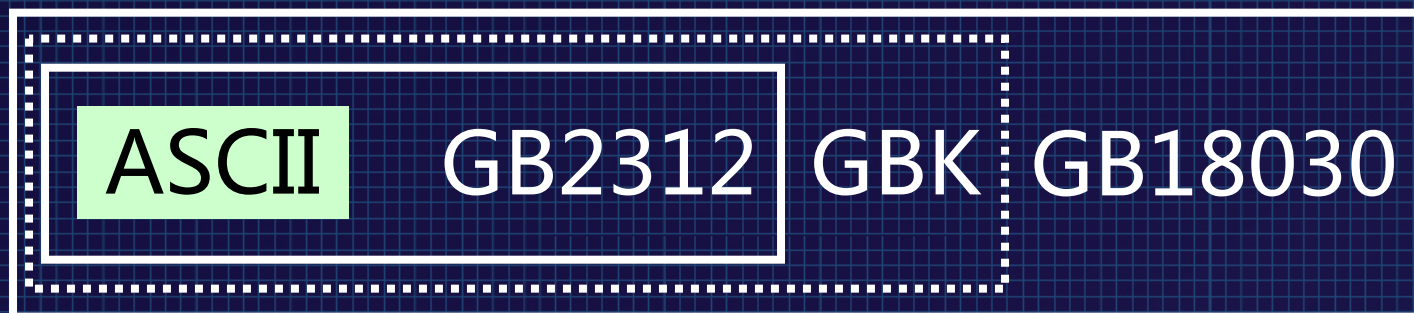
- 微软单方面对GB2312的扩充 – 双字节



- GBK的范围
 - GB2312中全部字符（兼容）
 - BIG5全部字符（不兼容）
 - 自定义区（windows.GBK ~= UNIX.GBK）

GB18030 – 多民族语言的补充

- 对GB2312的更大扩充 1、2、4字节



简/繁/民族文字

- GB18030的范围
 - 日文字符集
 - 韩文字符集
 - 简体/繁体中文+藏文/满文等

“字符集” 与 “编码”

- 字符集 一组具有共同特征抽象字符的集合
 - 英文字符集
 - ISO8859、CJK
 - 繁体字字符集 简体字字符集
 - 日文汉字字符集 日文假名字符集
- 编码 字符和二进制内码的对应码表
 - ASCII
 - ISO8859-1
 - GB2312

“字符集” 与 “编码”

- 同一种字符集的不同编码
 - CJK -> GB13000
 - CJK -> utf-8
- 同一种编码可以实现多个字符集
 - GB18030 -> 简体中文
 - GB18030 -> 繁体中文
 - GB18030 -> 英文字符集

- 编码之初
- ASCII
- 国际化
- **第二种国际化**
- HTTP之URL编码
- HTTP之BASE64编码
- 字体
- 编码之初

不兼容的悲剧

- 一段文本无法同时使用多种语言
- 一段文本无法同时使用简体和繁体

第二种国际化 – Unicode(2.0+)

- 万国码-Unicode

- 一种国际字符集，包含世界上绝大多数已知的字符集
- 定义一种编码“规则”，该集中每个字符唯一对应一个32位数值
- Unicode是包含字符集和码表的一个东西
- Unicode码表是具体编码的参照

又一个悲剧

- Unicode的不同编码对ASCII的兼容不统一
 - UTF-8兼容ASCII
 - UTF-16/32不兼容ASCII
- Unicode裸码需要四个字节描述一个字符

不是悲剧

- 多种语言共存
- 全球通用
 - Yahoo – Global
 - Taobao – China only

UTF (Unicode Translation Format)

- UTF - Unicode的存储
 - 裸存 将每个字符按照4个字存储
 - UTF-8 不同范围的字符使用不同长度的编码
 - UTF-16 始终使用2个字节存储一个字符
 - UTF-32 始终使用4个字节存储一个字符
 - UTF-32码表和Unicode码表是等价的
- UTF-8的中文字符存储占3个字

更多：

- Javascript内码采用unicode裸码
 - `alert('淘宝 '.length) == 2`
- 得到字符的10位unicode编码
 - `alert("淘".charCodeAt())`
- 得到字符的标准unicode编码(低位)
 - `alert(escape('淘').replace(/(u|%)/g,''))`
- 得到字符的UTF8编码
 - `alert(encodeURIComponent('淘').replace(/%/g,''))`

- 编码之初
- ASCII
- 国际化
- 第二种国际化
- **HTTP之URL编码**
- HTTP之BASE64编码
- 字体
- 编码之初

URL

- Uniform Resource Locator
- 统一资源定位符
- URL不受国别、语言差异的约束
- 是编码无关的
- RFC 1738
 - URL必须由英文字母、数字、和某些标点符号组成

非法的URL

- <http://www.baidu.com/s?wd=淘宝>



对URL进行编码

- 浏览器会对URL进行编码
 - ‘淘宝’ -> "%E6%B7%98%E5%AE%9D"
 - ‘淘宝’ -> "%CC%D4%B1%A6"(gb)
 - ‘淘宝’ -> "%6D%D8%5B%9D"(unicode)
- 影响URL编码的因素
 - 系统编码
 - 浏览器类型
 - web页面编码(form)

不同浏览器的地址栏URL编码

- Firefox
- IE

Firefox - 地址栏URL编码

- <http://www.baidu.com/s?wd=淘宝>

```
+ Transmission Control Protocol, Src Port: p
- Hypertext Transfer Protocol
  GET /s?wd=%CC%D4%B1%A6 HTTP/1.1\r\n
    Request Method: GET
    Request URI: /s?wd=%CC%D4%B1%A6
    Request Version: HTTP/1.1
    Host: www.baidu.com\r\n
    User-Agent: Mozilla/5.0 (windows; U; win
    Accept: text/html,application/xhtml+xml,
```

GB编码

IE - 地址栏URL编码

- <http://www.baidu.com/s?wd=淘宝>

```
Transmission Control Protocol, Src Port: 4030 (4030), Dst Port: http (80)
Hypertext Transfer Protocol
  GET /s?wd=%31%34%32%34%26%31%24%26 HTTP/1.1\r\n
    Request Method: GET
    Request URI: /s?wd=%31%34%32%34%26%31%24%26
    Request Version: HTTP/1.1
    Accept: */*\r\n
    Accept-Language: zh-cn\r\n
    UA-CPU: x86\r\n
    Accept-Encoding: gzip, deflate\r\n
    User-Agent: Mozilla/4.0 (compatible; MSIE 7.0; windows NT 5.1; CIBA; .
    Host: www.baidu.com\r\n
    Connection: keep-alive\r\n
000  00 21 27 3c 02 30 00 17 08 3d 9e be 08 00 45 00  .!'<.0.. .=....E.
010  01 66 2c 13 40 00 80 06 28 aa c0 a8 01 64 dc b5  .f,.@... (....d..
020  06 13 12 30 00 50 74 2b 3c 44 a8 d6 32 f8 50 18  ...0.Pt+ <D..2.P.
030  ff ff a6 2d 00 00 47 45 54 20 2f 73 3f 77 64 3d  ...-..GE T /s?wd=
040  cc d4 b1 a6 20 48 54 54 50 2f 31 2e 31 0d 0a 41  .... HTTP/1.1..A
050  63 63 65 70 74 3a 20 2a 2f 2a 0d 0a 41 63 63 65  ccept: */*..Acce
060  70 74 2d 4c 61 6e 67 75 61 67 65 3a 20 7a 68 2d  pt-Langu age: zh-
```

GB裸码

淘宝网

地址栏URL编码

- Firefox
 - 进行URL编码，编码方式和系统编码一致
- IE
 - 直接发送URL裸码，裸码编码和系统编码一致

Form表单中的URL编码

- GB系页面
- UTF-8 页面

GBK Form表单提交URL编码

- <http://www.baidu.com> 中搜索 “淘宝”

```
+ Transmission Control Protocol, Src Port: p
- Hypertext Transfer Protocol
  GET /s?wd=%CC%D4%B1%A6 HTTP/1.1\r\n
    Request Method: GET
    Request URI: /s?wd=%CC%D4%B1%A6
    Request Version: HTTP/1.1
    Host: www.baidu.com\r\n
    User-Agent: Mozilla/5.0 (windows; U; win
    Accept: text/html,application/xhtml+xml,
```

GB编码

UTF8 Form表单提交URL编码

- <http://www.google.com.hk> 搜索 “淘宝”

```
[-] Hypertext Transfer Protocol
[-] GET /search?hl=zh-CN&source=hp&q=%E6%B7%98%E5%AE%9D&btnG=Google+&
    Request Method: GET
    Request URI: /search?hl=zh-CN&source=hp&q=%E6%B7%98%E5%AE%9D&bt
    Request Version: HTTP/1.1
    Host: www.google.cn\r\n
    User-Agent: Mozilla/5.0 (windows; U; Windows NT 5.1; zh-CN; rv:1.
    Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*
```

UTF-8编码

Form表单提交URL编码

- GB系页面
 - URL编码和页面编码保持一致，gb系编码
- Utf-8页面
 - URL编码和页面编码保持一致，utf-8编码

不同应用对URL编码的接收处理

- 百度
 - <http://www.Baidu.com>
- 谷歌
 - <http://www.google.com>
- “淘宝” 的URL编码
 - Utf-8 "%E6%B7%98%E5%AE%9D"
 - GB系 "%CC%D4%B1%A6"

百度不支持UTF-8 URL编码

- <http://www.baidu.com/s?wd=%E6%B7%98%E5%AE%9D>



百度支持GB URL编码

- <http://www.baidu.com/s?wd=%CC%D4%B1%A6>



Google支持UTF8 URL编码

- <http://www.google.com.hk/search?q=%E6%B7%98%E5%AE%9D>



Google支持GB URL编码

- <http://www.google.com.hk/search?q=%CC%D4%B1%A6>



不同应用对URL编码的支持

- 百度
 - 支持gb系 URL编码
 - 不支持utf-8 URL编码
- 谷歌
 - 支持gb系 URL编码
 - 支持utf-8 URL编码
- 淘宝？

Ajax中的URL encode

- Javascript的encodeURIComponent始终采用utf-8编码

```
encodeURIComponent('淘宝')  
== "%E6%B7%98%E5%AE%9D"
```

- 编码之初
- ASCII
- 国际化
- 第二种国际化
- HTTP之URL编码
- **HTTP之BASE64编码**
- 字体
- 编码之初

BASE64

- BASE64和编码表不同，它是一种转换算法
- BASE64的目的
 - 可见明文传输
 - 回车、空格、二进制数据 -> 明文
 - 编码采用ASCII字符集
 - 防止编码范围溢出

Data URI (ie不支持)

- 在页面中嵌入二进制data

```

```

- 通过css嵌入

```
body{  
background:url(data:image/gif;base64,R0lGODlhAQAB  
AIAAAJRDvAAAACH5BAEAAAAALAAAAABAAEAAAIICRAEOw=  
=);}
```


- 编码之初
- ASCII
- 国际化
- 第二种国际化
- HTTP之URL编码
- HTTP之BASE64编码
- **字体**
- 编码之初

字体的显示 (windows系统)

GBK(环境相关)码表

操作系统

淘



CCD4

GBK编码



6DD8

unicode编码



字符映射表

淘

点阵

淘宝网

字符映射表



字体映射表基于unicode

淘宝网

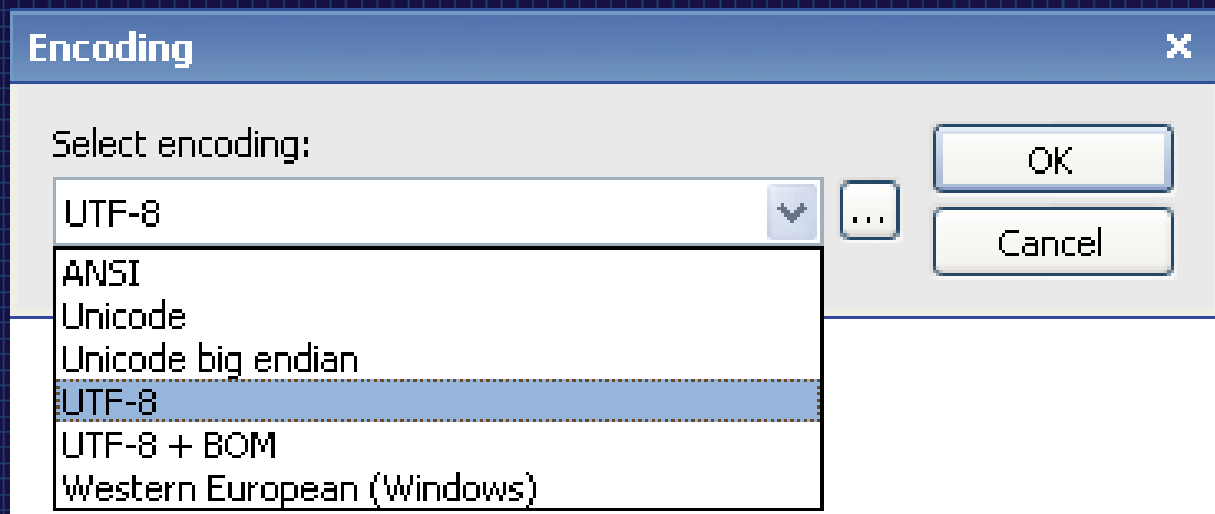
乱码的根源

- 编码表选择错误
- 字体错误

- 编码之初
- ASCII
- 国际化
- 第二种国际化
- HTTP之URL编码
- HTTP之BASE64编码
- 字体
- **编码之初**

表象

- OS中字符代码的表现形式



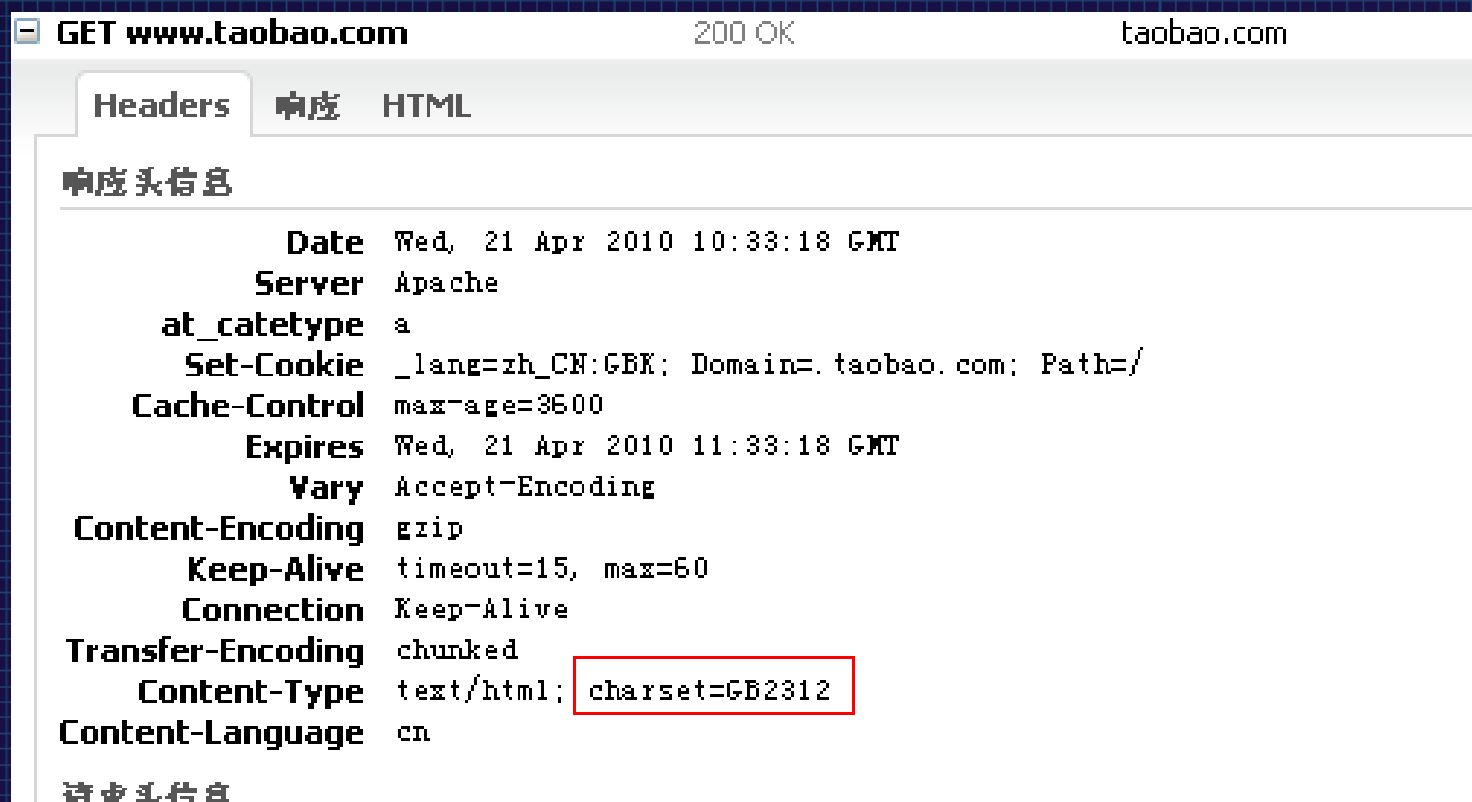
Windows中，文本编辑可选编码

表象

- ANSI：双字节编码的统称
 - 简体中文系统中，指代GB2312编码
 - 日文系统中，指代JIS编码
- 开发中如果页面存为GB，就选择ANSI
- Unicode和unicode big endian
 - 对于字节存储分别采用反序和正序

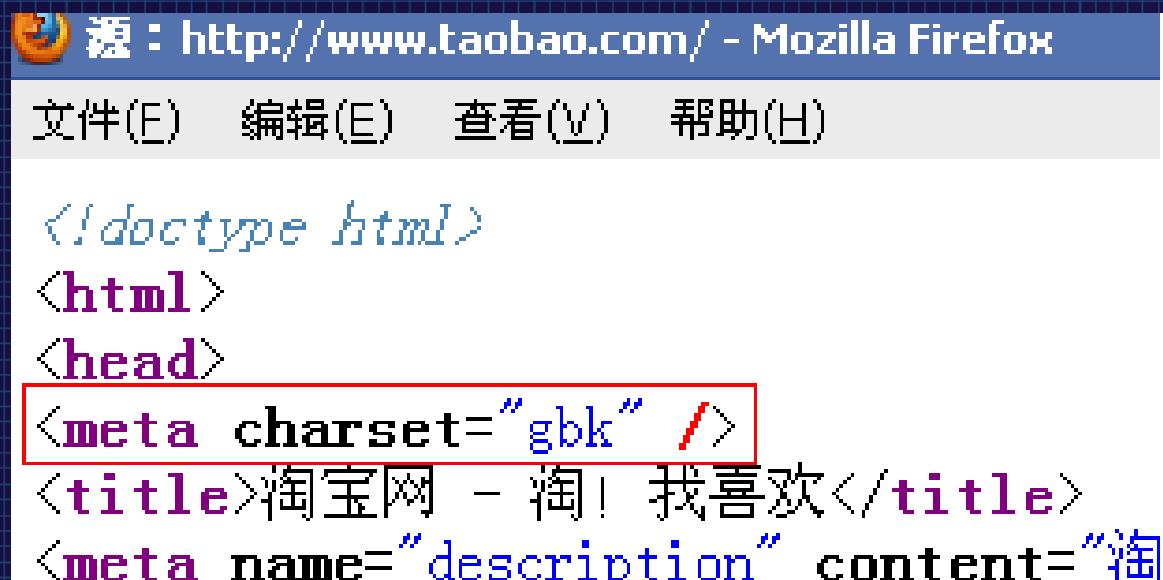
表象

- HTTP协议中的编码设置



表象

- 页面中的编码设置



```
源: http://www.taobao.com/ - Mozilla Firefox  
文件(E) 编辑(E) 查看(V) 帮助(H)  
<!doctype html>  
<html>  
<head>  
<meta charset="gbk" />  
<title>淘宝网 - 淘! 我喜欢</title>  
<meta name="description" content="淘
```

charset放置在title之前

表象

- 不同编码的外部脚本引入页面

```
<script charset="gb18030" type="text/javascript" src="http://a.tbcdn.cn/
```


本质

- 一码是一码

参考

- BIG5码表
<http://www.geo.ntnu.edu.tw/faculty/hchou/class/ntptc/gis/code&translate.htm>
- GB2312码表
<http://www.knowsky.com/resource/gb2312tbl.htm>
- Gbk码表
<http://58.248.189.53/SchoolWeb/hzdwzx/xxzy/xxzy-kj/xxzy-xx/xxzy-xx1/xxzy-xx1-1/HAIZI/GBK2.htm>
- Unicode的中日韩部分码表
<http://www.chi2ko.com/tool/CJK.htm>
- Iso8859
http://en.wikipedia.org/wiki/ISO/IEC_8859

Q&A