Punycode

山本和彦
(株)インターネットイニシアティブ
kazu@iij.ad.jp

Punycode とは？

- ピュニコード
  - puny は「ちっぽけな」の意味
- Unicode である IDN を
  ASCII の図形文字へ変換する
  - IDN = Internationalized Domain Name
- IETF WG は ACE の中から Punycode を採択
  - JPNIC (現 JPRS) が提案していたのは RACE
  - ACE = ASCII Compatible Encoding
  - RACE = Row-based ASCII Compatible Encoding
- RFC 3492
  - March 2003

例題

- 日本語 → xn--wgv71a119e
  - "xn--" が Punycode の始まり
  - Punycode の部分は、"-", 0〜9, A〜Z, a〜z
- 3年B組 → xn--3b-w52dz04i
  - ASCII の図形文字が混ざると "-" が区切りとなる
- 符号化と復号化を体験したい人は
  - http://punycode.jp/

Punycode の目標

- 一対一の対応
  - IDN の文字列 ←→ Punycode で符号化された文字列
- 符号化の結果が短いこと
  - ドメイン・ラベルは 63文字まで
- 簡潔なこと
  - 本当？
- Unicode の基本ラテン文字は、それ自身で表現
  - u+0000 ～ u+007F → 0x00～0x7F

符号化の概要

- 例
  - 3年B組金八
- Unicode
  - u+0033 u+5E74 u+0042 u+7D44 u+91D1 u+516B
- 基本ラテン文字を前へ
  - u+0033 u+0042 u+5E74u+7D44 u+91D1 u+516B
  - "3b-" ＋ u+5E74 u+7D44 u+91D1 u+516B
- 小さい順に並べ替え
  - "3b-" ＋ u+516B u+5E74 u+7D44 u+91D1
- 差分の計算
  - "3b-" ＋ 20843 3337 7888 5261
- ASCII へ
  - "3b-wz4c970c5k0b50za"

(注意) 正確ではありません

疑問

- 元の位置は、どうやって知るのか？
  - 基本ラテン文字を前に送っている
  - 並べ替えている
- 区切りは、どこか？
  - "wz4c970c5k0b50za" のどこが 「年」か？

元の位置

- 差分に元の位置を加える
  - 差分 × 元の文字列の大きさ ＋ 元の位置
- 考察
  - 元の文字列の大きさ分ビット数を使うより、
    元の文字列の大きさを掛ける方が小さい？
  - 例) 6 で 3 ビット消費するより、6倍する方が小さい？

(注意) 正確ではありません

一般可変長整数

- 普通の整数
  - 例) ... 152437
  - 7 or 37 or 437 or 2437... ？
- 一般可変長整数では閾値を設ける
  - 最上位の桁のみが閾値より小さくなる
  - よって区切りが分かる
- 一般可変長整数の例
  - 一般可変長整数）... 1 5 2 4 3 7
  - 閾値）　　... 5 5 2 5 3 2
  - 152 と 437

桁の意味

- 一般式
  - $x_n \times w_n$
- 普通の整数
  - $w_n = 10^n$
  - $x_n \times 10^n$
- 一般可変長整数
  - $w_0 = 1$
  - $w_n = w_{n-1} \times (10 - t_n)$, *for* $n > 0$
  - t_n は閾値
- 例 437 [532]
  - $w_0 = 1$
  - $w_1 = 1 \times (10 - 2) = 8$
  - $w_2 = 8 \times (10 - 3) = 56$
  - $4 \times 56 + 3 \times 8 + 7 \times 1 = 255$

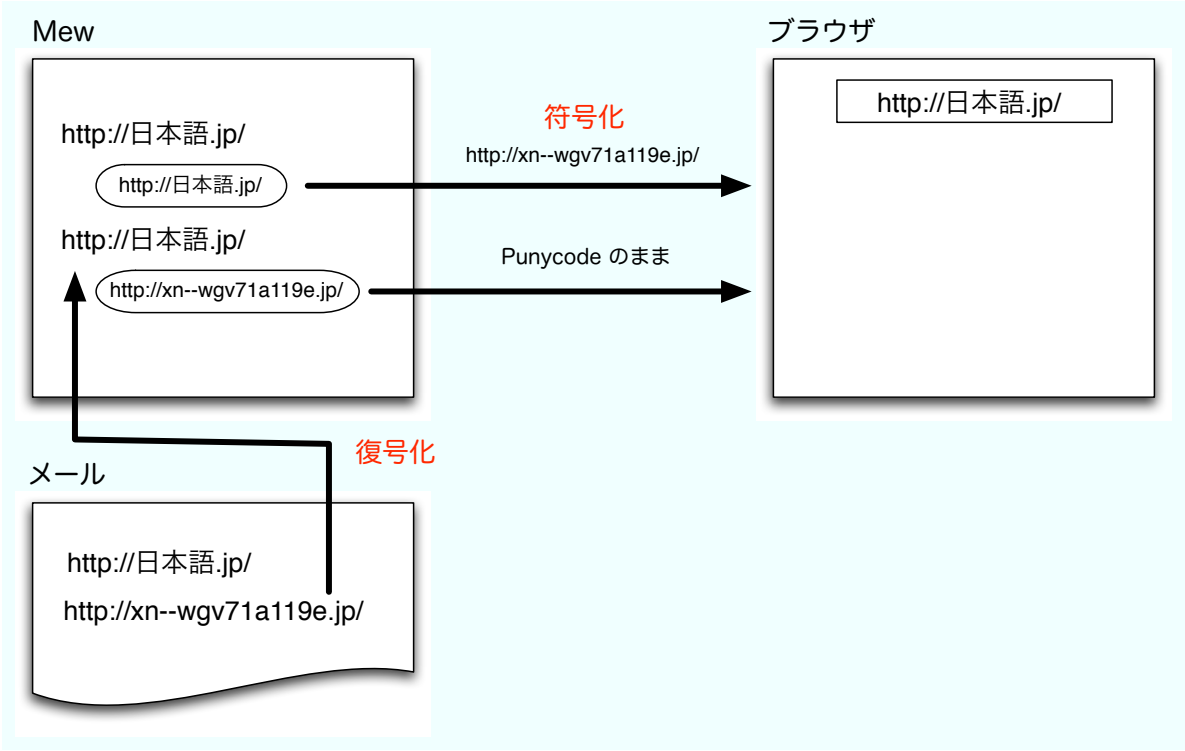
閾値

- 閾値を決める式
  - $t_n = 10 \times (n + 1) - bias$
- さらに *bias* を決める式がある...

Mew での実装

- まともな Punycode ライブラリがない
  - punycode.el は外部コマンドを呼び出す
  - → 自前で作る
- ブラウザへの URL 受け渡し
  - Mac OS X ではコマンド "open" を使う
  - "open" は、引数の URL に Unicode を許さない
  - punycode のままで渡す必要がある
- メールのヘッダ
  - とりあえず放置
- メールの本文
  - ISO-2022-JP などで書かれた IDN の URL
  - Punycode の URL

Mew での実装


Mew
ブラウザ
http://日本語.jp/
http://日本語.jp/
符号化
http://xn--wgv71a119e.jp/
http://日本語.jp/
http://日本語.jp/
Punycode のまま
http://xn--wgv71a119e.jp/
復号化
メール
http://日本語.jp/
http://xn--wgv71a119e.jp/