I set di caratteri

Fabio Vitali

5 novembre 1999

Introduzione

Qui esaminiamo in breve:

- Il problema della codifica dei caratteri
- ASCII (7 bit ed esteso)
- ISO/IEC 10646 e UNICODE
- UCS e UTF

I set di caratteri

- La globalizzazione di Internet ha proposto il problema di rendere correttamente gli alfabeti di migliaia di lingue nel mondo.
- Il problema non si pone per i protocolli, che trattano byte interpretati da applicazioni, anche se “per caso” sono significativi per persone di lingua inglese quando scritti in US-ASCII
- Il problema si pone per il contenuto dei protocolli, in quanto deve essere evidente e non ambiguo il criterio di associazione di un blocco di bit ad un carattere di un alfabeto.

I caratteri (1)

Il carattere è l'entità atomica di un testo scritto in una lingua umana.

In alfabeti diversi i caratteri hanno particolarità diverse:

- Negli alfabeti di derivazione greca (greco, latino e cirillico), esiste la distinzione tra maiuscole e minuscole, ignota altrove
- Negli alfabeti di derivazione latina si sono inventati segni particolari sulle lettere per soddisfare le esigenze dei singoli alfabeti (accenti, segni diacritici, ecc.).
- In ebraico, le vocali sono modificatori di lettere di consonanti
- In arabo, la giustapposizione di lettere diverse nella parola provoca una differenziazione della forma delle lettere stesse.
- In cinese, è possibile creare nuovi caratteri come composizione di altri caratteri esistenti.

I caratteri (2)

Lingue diverse associano ai caratteri ruoli diversi: rappresentano di volta in volta suoni, sillabe, intere parole.

Esistono tre aspetti di un carattere:

- La sua natura (di difficile attribuzione: a e à sono la stessa lettera?)
- La sua forma, o glifo (con ambiguità: P ha un suono negli alfabeti latini, e un altro negli alfabeti greci e cirillici; inoltre i font creano forme anche molto diverse per le stesse lettere).
- Il suo codice numerico: in base ad una tabella piuttosto che un'altra, lettere diverse, di alfabeti diversi, hanno lo stesso codice numerico, o la stessa lettera ha codici diversi.

ASCII, EBCDIC e ISO 646

- ASCII (American Standard Code for Information Interchange) è uno standard ANSI (X3.4 - 1968) che definisce valori per 128 caratteri, di cui 33 (0-31 e 127) non stampabili. Nello standard originale il primo bit non era significativo e ne sono state fatte estensioni non standard e proprietarie sui codici alti.
- EBCDIC (Extended Binary Characters for Digital Interchange Code) è il codice caratteri a 8 bit usato da IBM nei suoi mainframe.
- ISO 646 definisce un sottoinsieme di 83 caratteri comune tra ASCII, varie famiglie di EBCDIC, e modifiche nazionali di ASCII a 7 bit (tranne ! e ?, che sono incompatibilmente diversi).

ISO 8859/1 (ISO Latin 1)

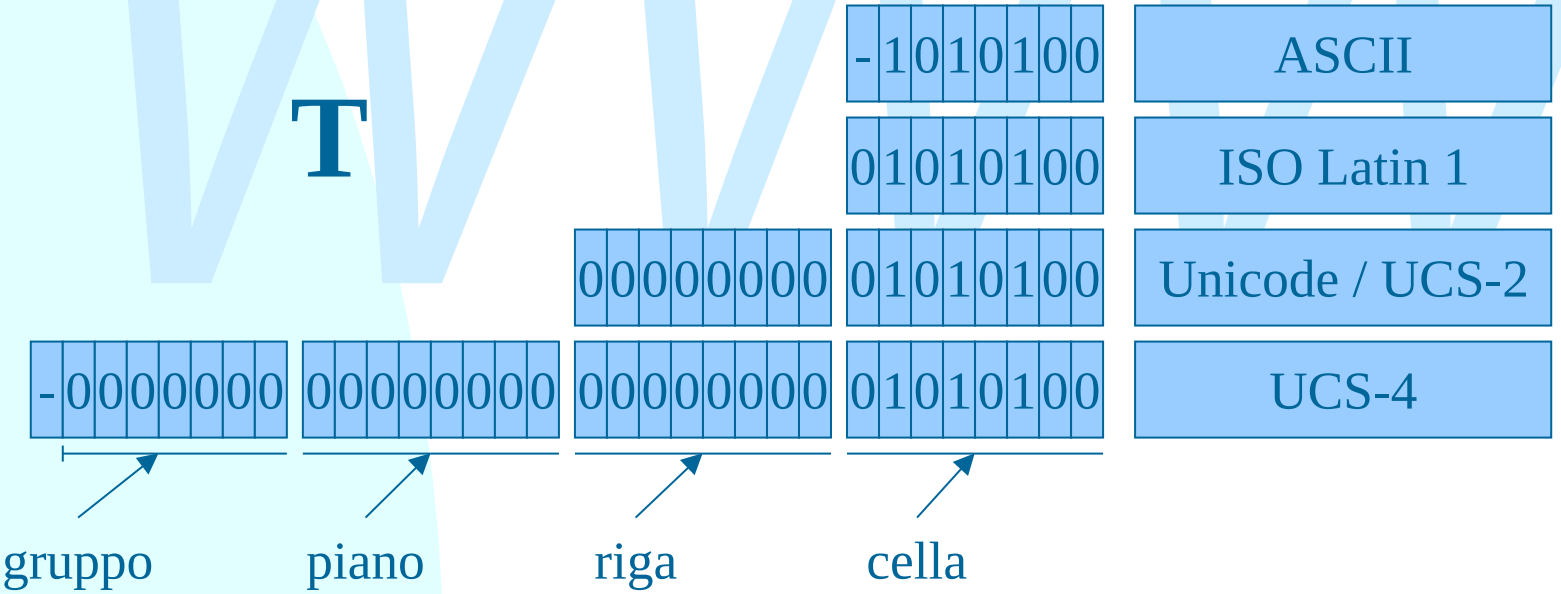
- Estensioni di ASCII sono state fatte per utilizzare il primo bit e accedere a tutti i 256 caratteri. Nessuna di queste è standard
- ISO 8859/1 (ISO Latin 1) è l'unica estensione standard e comprende un certo numero di caratteri degli alfabeti europei come accenti, ecc.
- ISO Latin 1 è usato automaticamente da HTTP e qualche sistema operativo.
- Ovviamente ISO Latin 1 è compatibile all'indietro con ASCII, di cui è un estensione per i soli caratteri >127.

Unicode e ISO/IEC 10646 (1)

- Esistono dozzine di codici a 8 bit per alfabeti non latini (e.g., cirillico, greco e giapponese semplificato) e molti codici a 16 bit per linguaggi orientali (giapponese e cinese).
- Il compito di creare uno standard unico è stato affrontato indipendentemente da due commissioni di standard, Unicode e ISO/IEC 10646.
- Unicode usa 2 byte per ogni carattere, per un totale di 65536 caratteri. Questo basta per coprire la maggioranza degli alfabeti, ma non il cinese.
- ISO 10646 utilizza uno schema a lunghezza variabile fino a quattro byte, con 2 miliardi di combinazioni. Anche Unicode 2.0 ha introdotto schemi a lunghezza variabile.

Unicode e ISO/IEC 10646 (2)

- ISO 10464 è composto di vari schemi di codifica. Il più semplice utilizza un numero fisso di byte:
- UCS-2 è uno schema a due byte, ed è identico a Unicode (questo non è garantito in futuro). E' un'estensione di ISO Latin 1.
- UCS-4 è uno schema a 31 bit in 4 byte, estensione di UCS-2. E' diviso in gruppi, piani, righe e celle.

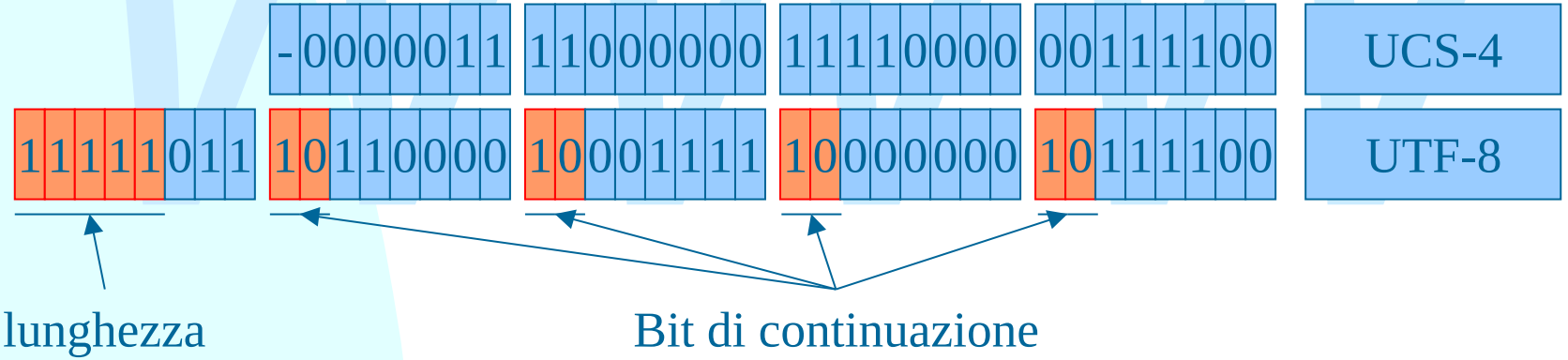


Unicode e ISO/IEC 10646 (3)

- In UCS-4 esistono dunque 32768 piani di 65536 caratteri ciascuno. Il primo piano, o piano 0, è noto come BMP (Basic Multilingual Plane) ed è ovviamente equivalente a UCS-2 e a Unicode.
- Tuttavia, nella maggior parte dei casi i testi scritti utilizzeranno soltanto uno degli alfabeti del mondo. In questo caso, sono necessari soltanto una minima parte dei caratteri di UCS-4.
- Inoltre, la maggior parte degli alfabeti sta nel BMP, e la maggior parte dei documenti sono scritti in ASCII.
- E’ dunque uno spreco utilizzare quattro byte per ogni carattere in questo caso.

Unicode e ISO/IEC 10646 (4)

- Sono quindi stati sviluppati due schemi di compressione per utilizzare un byte o due byte per UCS-4. Questi schemi a lunghezza variabile si chiamano UTF (UCS Transformation Format).
- UTF-8 utilizza 1 byte per caratteri ASCII a 7 bit, e da due a sei byte per i caratteri estesi.
- Il primo bit deve essere a uno. All'inizio viene scritta la lunghezza in byte dell'intero carattere, poi il carattere viene scritto in blocchi di sei bit preceduti dal codice '10' per indicare a prosecuzione del carattere.



- Unicode è equivalente a UCS-2, ma ha un meccanismo a lunghezza variabile chiamato UTF-16, che permette di accedere ai primi 16 piani di UCS-4.

Conclusioni

Qui abbiamo parlato di set di caratteri

- A lunghezza fissa, 7, 8 bit (ASCII, EBCDIC, ISO Latin 1)
- A lunghezza fissa, 16, 31 bit (UCS-2, UCS-4)
- A lunghezza variabile, 2-6 * 8 bit (UTF-8, UTF-16)

Riferimenti

N. Bradley, The XML companion, Addison Wesley, 1998, cap. 13.

K. Simonsen, Character Mnemonics & Character Sets, RFC 1345, IETF, June 1992