特集 SPECIAL FEATURE 1

文字コードの自動判別機能実装

文字コードの特徴を知ってコード自動検出機能を実装する

大澤 文孝 OSAWA, Fumitaka

Level

1 2 3 4 5

Technology Tools

- [x] Visual Basic
- [ ] Visual C#
- [ ] Visual C++
- [ ] SQL Server
- [ ] Oracle
- [ ] Access
- [ ] Excel
- [ ] ASP.NET
- [ ] Other:

Samples

この記事で取り上げたソースコードおよびサンプルプログラムは、
http://www.shoeisha.com/mag/windev/
からダウンロード可能です。

はじめに

文字コードを変換する場合、「変換元の文字コードがわからない」ということがあります。

そのような場合には、バイトの並びを調べて、変換元となる文字コードを推測する処理が必要となります。

「JIS」「シフトJIS」「EUC」「Unicode」「UTF-8」の各文字コードには、それぞれバイト列の並びに特徴があり、変換元の文字列が極端に短くなければ、その特徴を抽出して、文字コードを判定できます。

本稿では、それぞれの文字コードの特徴と、その特徴から文字コードを自動判定する方法を説明します。

文字コードの基礎

文字コードとは、「字形に対して、どのような数値（コード）を割り当てるのか」という規則のことです。

文字コードは、俗に「半角文字」と呼ばれる「1バイトコード」と、「全角文字」と呼ばれる「2バイトコード」に分類できます。

半角文字の基本となるアスキーコード

このうち半角文字部分は、「アスキーコード」と呼ばれ、英数字や半角カナが定義されています（図1）[注1]。

アスキーコード部分は、「JIS」「シフトJIS」「EUC」「Unicode」「UTF-8」のどれでも同じです。

ただしUnicodeの場合には、すべてを2バイトとして表記します。たとえば、アスキーコード「&H41」の「A」という文字は、「&H41」「&H00」というバイトの並びになります[注2]。

図1に示したように、アスキーコードでは、「&H00～&H1F」と「&H7F」は、

注1）「&H80～&HFF」までの範囲は、英語圏では、半角カナではなくて、ラテン文字や記号などが割り当てられています。図1に示しているのは、JIS規格として定められた「JIS X 0201」というコード表です。

注2）Unicodeでは、文字「A」は「&H0041」です。Windowsでは、「下位バイト／上位バイト」の順で表記するリトルエンディアンが一般的なので、バイト順は、「&H41」「&H00」となります。「上位バイト／下位バイト」の順で表記するビッグエンディアンの場合には「&H00」「&H41」の並びになります。

特殊な制御コードとして使われます。これらの範囲内には、タブや改行などのコードが含まれています。

全角文字の基本となる区点コード

日本では、日本規格協会（JSA）が、文字コードをJIS規格として定めています。その基本となるのが、「区点コード」です。

区点コードでは、文字を94×94のマトリックスで定義します。区点コードは、「1区1点」から「94区94点」まで定義されます（区と点は、それぞれ「1」から始まります。「0」からではありません）。

区点コードは、文字の種類や読みなどを基準として並べたもので、たとえば「漢」という文字は、区点コード表より、「20区33点」となります（図2）[注3]。

文字コードには、「JIS」「シフトJIS」「EUC」「UTF-8」「Unicode」といったように、いくつかの種類がありますが、どれも区点コードをベースとしたものになっています。

それぞれの文字コードの特徴

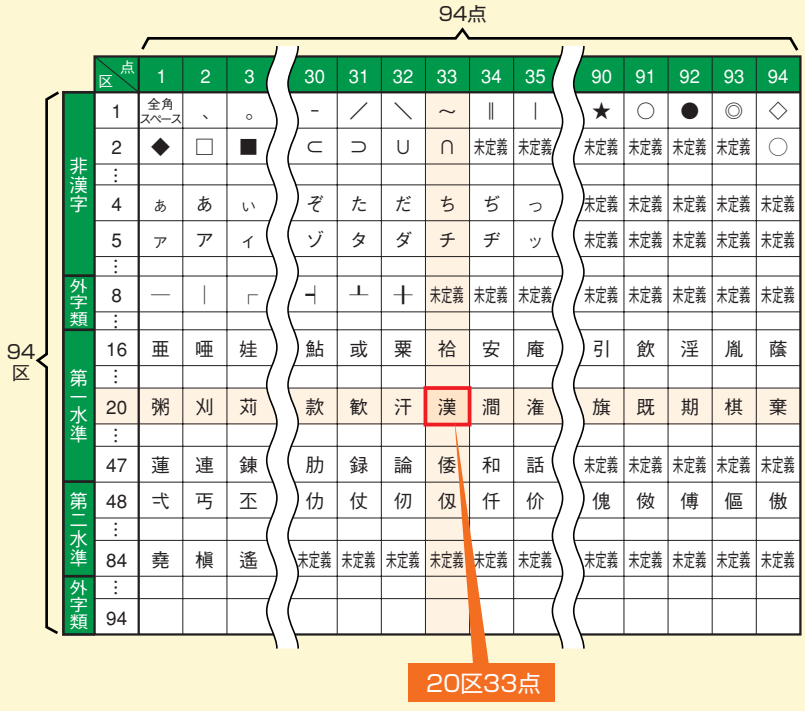
では次に、それぞれの文字コードは、区点コードをどのように変化させているのか、その特徴を見ていきましょう。

図1：アスキーコード

上位／下位	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	DLE	スペース	0	@	P	`	p	未定義	未定義	未定義	－	タ	ミ	未定義	未定義
1	SOH	DC1	!	1	A	Q	a	q	未定義	未定義	。	ア	チ	ム	未定義	未定義
2	STX	DC2	"	2	B	R	b	r	未定義	未定義	「	イ	ツ	メ	未定義	未定義
3	ETX	DC3	#	3	C	S	c	s	未定義	未定義	」	ウ	テ	モ	未定義	未定義
4	EOT	DC4	$	4	D	T	d	t	未定義	未定義	、	エ	ト	ヤ	未定義	未定義
5	ENQ	NAK	%	5	E	U	e	u	未定義	未定義	・	オ	ナ	ユ	未定義	未定義
6	ACK	SYN	&	6	F	V	f	v	未定義	未定義	ヲ	カ	ニ	ヨ	未定義	未定義
7	BEL	ETB	'	7	G	W	g	w	未定義	未定義	ァ	キ	ヌ	ラ	未定義	未定義
8	BS	CAN	(	8	H	X	h	x	未定義	未定義	ィ	ク	ネ	リ	未定義	未定義
9	HT	EM	)	9	I	Y	i	y	未定義	未定義	ゥ	ケ	ノ	ル	未定義	未定義
A	LF	SUB	*	:	J	Z	j	z	未定義	未定義	ェ	コ	ハ	レ	未定義	未定義
B	VT	ESC	+	;	K	[	k	{	未定義	未定義	ォ	サ	ヒ	ロ	未定義	未定義
C	FE	FS	,	<	L	¥	l	¦	未定義	未定義	ャ	シ	フ	ワ	未定義	未定義
D	CR	GS	-	=	M	]	m	}	未定義	未定義	ュ	ス	ヘ	ン	未定義	未定義
E	SO	RS	.	>	N	^	n	~	未定義	未定義	ョ	セ	ホ	゛	未定義	未定義
F	SI	US	/	?	O	_	o	DEL	未定義	未定義	ッ	ソ	マ	゜	未定義	未定義

制御コード　　シフトJISコードの第1バイト目

図2：区点コード



注3）文字コードは、16進数で示されることがほとんどですが、区点コードについては、10進数で示されるのが一般的です。ここで例に挙げている「20区33点」は10進数です。