TẠP CHÍ KHOA HỌC TÍNH TOÁN VÀ ĐIỀU KHIỂN TẬP III (1987), SỐ 2 (27—30)

MỘT PHƯƠNG ÁN THIẾT KẾ CÀI ĐẶT BỘ CHỮ VIỆT TRÊN MÁY VI TÍNH

ĐÀO HỮU CHÍ, LÊ MẠNH

Trong hội thảo « Các hệ xử lý văn bản tiếng Việt trên máy vi tính tổ chức vào tháng 1 1987 có nhiều nhu cầu của những người làm công tác tin học trong và ngoài nước đối với việc cài đặt chữ Việt trên các máy vi tính. Trong hội thảo, nhiều kết quả cài đặt đã được giới thiệu và phân tích một cách sâu sắc. Trong bài báo này, chúng tôi muốn thảo luận về một số điểm tồn tại và trên cơ sở đó đề xuất một hướng giải quyết nhằm góp phần nhanh chóng đi đến một giải pháp thống nhất.

Chúng tôi nhận thấy rằng cần phải xem xét lại một cách nghiêm túc những tiêu chuẩn đề ra cho một hệ xử lý và soạn thảo văn bản chữ Việt trên máy tính bởi vì chỉ có trên cơ sở các tiêu chuẩn này thì chúng ta mới có hướng lựa chọn một giải pháp hợp lý. Xin đưa ra một ví dụ, chúng ta đưa tiêu chuẩn cài đặt bộ mã chữ Việt phải đảm bảo sự hoạt động bình thường của hầu hết các chương trình tiện ích (utility) hiện có trong máy vi tính. Vậy « hầu hết » đó là bao nhiêu phần trăm? Có người nghĩ rằng có thể sử dụng 128 mã sau của bảng mã ASCII (American Standard Code for Information Interchange) để mã hóa các ký tự chữ Việt, nhưng khi sử dụng bộ mã như vậy, có nhiều chương trình tiện ích không chạy được như Wordstar...). Do đó, trong bối cảnh bị ràng buộc bởi thiết bị máy móc và các bộ chương trình tiện ích nhập từ nước ngoài, chúng ta khó có thể hy vọng một giải pháp « triệt để ». Cũng có khó khăn tương tự như thế khi thể hiện trên màn hình. Có người mong muốn giải quyết triệt để hơn, bằng cách thể hiện cả chữ thường lẫn chữ in hoa đều có dấu. Điều này không thể thực hiện được vì bộ chữ đó vượt quá cả 128 ký tự sau của bảng ASCII. Chúng tôi đã cài đặt việc thể hiện trên màn hình chỉ chữ thường có dấu, chữ in hoa không dấu, nhưng khi in ra trên máy in vẫn thể hiện dấu cả chữ hoa lẫn chữ thường. Việc cài đặt bộ mã chữ Việt nằm trong 128 mã sau của bảng ASCII ở mức hệ thống chủ yếu phục vụ các nơi sử dụng. Chúng tôi đã cài đặt các hệ phục vụ soạn thảo văn bản chữ Việt, nhưng nếu cần xử lý văn bản đó (sắp xếp, so sánh v.v...) thì phải có hệ chương trình riêng. Qua kinh nghiệm cài đặt và khi đưa cho người sử dụng bộ chương trình chữ Việt như vậy chúng tôi muốn thống nhất một vài quan điểm và từ các quan điểm đó sẽ dễ dàng trong việc so sánh, lựa chọn các giải pháp khác nhau [1, 2].

Khi xem xét chi tiết các khía cạnh cài đặt cụ thể, chúng ta đều biết rằng bất kỳ một giải pháp nào cũng đều giải quyết 3 khâu cơ bản:

— Mã hóa các ký tự chữ Việt ở bộ nhớ trong (RAM).

— Nhập văn bản chữ Việt và ra văn bản thể hiện lên màn hình và máy in.

— Lưu trữ văn bản đó trên các thiết bị nhớ ngoài.

Mặc dù 3 khâu này mang tính chất độc lập tương đối, nhưng có mối quan hệ chặt chẽ với nhau. Đảm bảo được các chức năng của hệ thống cũ là một vấn đề nan giải. Mức khó khăn ở đây tùy thuộc vào việc lưu trữ bên trong RAM như thế nào. Nhìn chung, chúng ta có thể phân ra hai phương pháp lưu trữ:

— Phương pháp thứ nhất là mỗi ký tự biểu diễn bên ngoài (ở trên màn hình) đều thể hiện bằng một mã tương ứng ở bên trong RAM.

— Phương pháp thứ hai là lưu trữ ký tự chữ Việt trong RAM khác (nhiều hơn) so với việc thể hiện trên màn hình.

Phương pháp thứ nhất đã được nhiều nơi cài đặt và đưa vào ứng dụng [1, 2]. Việc sử dụng các bộ chương trình đó tại nhiều cơ sở sản xuất, ứng dụng tin học và quản lý kinh tế v.v... có một vài kết quả khả quan. Các phần mềm tiện ích quan trọng như DBASE II+III, FRAMEWORK, các chương trình dịch ngôn ngữ bậc cao BASIC, FORTRAN 77, TURBO—PASCAL,

C-LANGUAGE đều có thể dùng bộ chữ Việt đó.

Ưu điểm của phương pháp này là:

1. Bộ mã tạo ra đơn giản dễ sử dụng.
2. Việc tạo ra hoặc sửa đổi các chương trình driver dễ dàng, cả trong hệ điều hành.
3. Việc tạo ra, việc sửa đổi bàn phím cho việc nạp văn bản cho bộ chữ Việt đơn giản.
4. Việc cập nhật, xen, xóa và lưu trữ văn bản gọn gàng và dễ dàng (cả trong mức hệ thống lẫn trong các chương trình tiện ích).

Tuy nhiên phương pháp này còn có một số nhược điểm như sau:

1. Không đảm bảo thứ tự sắp xếp theo từ vựng tiếng Việt (cần phải thêm các đơn thể sắp xếp đặc biệt).

2. Sử dụng quá nhiều mã ở 128 mã ngoài bảng ASCII dẫn đến tình trạng nhiều chương trình tiện ích không dùng được. Ngay cả dùng hết 128 mã ngoài bảng ASCII cũng không có khả năng thể hiện dấu cho cả chữ hoa lẫn chữ thường.

Để khắc phục nhược điểm của phương pháp thứ nhất, chúng tôi sẽ bàn đến mặt logic và đưa ra một vài gợi ý trong cách cài đặt phương pháp thứ hai. Có một số nơi dùng mã điện tín trong bưu điện để lưu trữ bên trong RAM [3]. Điều này đã bị bác bỏ, vì hai nguyên nhân sau:

1. Không có khả năng ánh xạ 1—1 trong việc so sánh và thể hiện từ RAM ra nghĩa cụ thể của người sử dụng muốn có.

2. Việc sắp xếp, xen, xóa v.v... không thực hiện bình thường ở các chương trình tiện ích và cả trong hệ thống.

Sau khi tham khảo các tài liệu chúng tôi có trong tay về các hệ điều hành của Microsoft như Avanced MS—DDS và DOS 281 dùng cho các máy vi tính IBM—PC/XT hoặc AT có các bộ xử lý trung tâm 80286 hoặc 80386, chúng tôi đưa ra một phương pháp thiết kế và gợi ra một vài giải pháp trong khi cài đặt một bộ soạn thảo và xử lý chữ Việt khác với các bộ đã có trong nước ta.

Tiêu chuẩn chủ yếu của chúng tôi là:

1. Bộ chữ Việt có khả năng tối đa nhúng vào các sản phẩm phần mềm (các chương trình tiện ích, các hệ điều hành...) hiện có hoặc sẽ có trên thế giới.

2. Các đơn thể tìm kiếm, so sánh, cập nhật, sửa, xen... có thể dùng ngay các chương trình đã có trong DOS hoặc trong các chương trình tiện ích.

3. Bộ mã chữ Việt mới này ít thay đổi so với bộ chữ Anh có trong máy.

4. Không động chạm đến 128 bytes cao ngoài bảng ASCII.

Giải pháp của chúng tôi đặt ra là dùng 10 mã tự do trong bảng mã ASCII để thể hiện các dấu và ký tự đặc biệt của chữ Việt. Ngoài ra chúng tôi còn dùng thêm một nguyên tắc ghép hai chữ thành một từ cho đúng với ngữ pháp của tiếng Việt. 10 mã tự do trong bảng ASCII thường được dùng cho các nước có chữ khác Anh (như chữ Đức, Italia, Tây ban nha v.v..) Với 10 mã tự do đó chúng tôi sẽ thể hiện 5 dấu bằng, sắc, hỏi, ngã, nặng) 3 dấu ghép với nguyên âm e, a, u, o trở thành nguyên âm khác trong chữ Việt (cho chữ ê, ă, â, ô, ơ, ư), dấu gạch trên cho chữ đ và một từ nối cho ghép hai từ thành một từ.

MÃ CỦA 10 KÝ TỰ TỰ DO TRONG BẢNG MÃ ASCII *Bảng 1*

HEX	DEC	Mã ASCII chữ Anh	Mã ASCII chữ Đức	Mã ASCII chữ Tây ban nha		Chữ Việt Nam
5B	91	[	Ä	↑		Dấu huyền
5C	92	\	Ö	Ñ		Dấu sắc
5D	93	]	Ü	ċ		Dấu hỏi
5E	94	^	^	^		Dấu ngã
5F	95	_	_	-		Dấu nặng
60	96	`	`	`		Nối từ
7B	123	{	ä	¨	(*)	Dấu chữ đ
7C	124	|	ö	ñ		Dấu chữ ă
7D	125	}	ü	}		Dấu chữ â ô ê
7E	126	~	β	~		Dấu chữ ơ ư

Việc bố trí bên trong RAM theo nguyên tắc sau:

— Các dấu trong chữ Việt luôn luôn xếp cuối một từ.

— Các dấu đặc biệt sau dấu (*) ở bảng 1 luôn luôn xếp ngay chữ nó cần thể hiện, ví dụ đ xếp d—, ă xếp a ◡, v.v..

— Ký tự nối từ khi có sẽ được xử lý đưa hết các dấu của từ trên xuống dưới, ví dụ lúc lắc trong RAM khi gặp dấu nối từ (—) sẽ chuyển như sau: lục lạc

Việc tìm nguyên âm đầu tiên để đưa dấu vào khi thể hiện lên màn hình hoặc đưa ra máy in đã được nêu trong phụ lục A [4].

Chúng tôi cũng đưa ra một vài gợi ý khi cài đặt theo nguyên tắc này như sau:

— Việc lưu trữ trong RAM không đúng với việc thể hiện trên màn hình sẽ ảnh hưởng đến các lệnh cập nhật, xen và xóa các chương trình tiện ích. Nhưng với sự tồn tại bộ nhớ tại chỗ (Memory Resident) ở những địa chỉ thấp trong RAM, chúng ta sẽ đưa toàn bộ một Record ra Memory Resident (coi đây như là buffer của màn hình) để thao tác trên đó. Sau khi thao tác trên đó, từ Memory Resident sẽ đưa trở lại RAM. Chúng ta sẽ dùng phần Side—Kick (với 18 nhịp đồng hồ) để đưa lại những phần đã được thao tác lên màn hình. Quá trình này với tốc độ CPU 8 đến 12MHz có khả năng thực hiện «tức khắc» không làm ảnh hưởng đến thói quen của người sử dụng.

— Việc «vẽ» một chữ Việt trong một từ trong RAM lên màn hình hoặc máy in có những dấu đặc biệt, chúng ta dùng các nhịp trong phần Side—Kick sẽ không làm ảnh hưởng đến quá trình xử lý trong việc truy nhập vào các phần của hệ điều hành đưa ra máy.

— Khi đưa các ký tự bố trí trong RAM hoặc ở các thiết bị nhớ ngoài không đúng với thể hiện logic của từ đó lên màn hình, chúng ta cần viết thêm bộ tiền xử lý. Bộ tiền xử lý này sẽ sắp xếp lại, bố trí đưa lên màn hình đúng theo yêu cầu. Thường bộ xử lý này sẽ viết ngay vào phần ECHO trong DOS. Nhưng cũng có các chương trình tiện ích không sử dụng phần ECHO trong DOS thì phải sửa ngay trong chức năng Interrupt 10 và 17 trong System—calls. Qua bộ tiền xử lý này chúng ta sẽ đưa một từ Việt từ RAM hoặc bộ nhớ ngoài ra màn hình đúng theo yêu cầu.

— Trong bảng 2, chúng ta nhận thấy các dấu của chữ Việt giữa các chữ hoa và chữ hường, do đó trong DOS ta thêm đơn thể cho phép biến tất cả chữ hoa sang chữ nhỏ (như hàm UPPER trong BASIC). Điều này cho phép chúng ta sắp xếp mới đúng được.

Trên đây là một số gợi ý để xây dựng, thiết kế và cài đặt một hệ xử lý và soạn thảo cho chữ Việt trên những máy vi tính 16 bit. Chắc chắn phải qua thử nghiệm cài đặt và đưa thử vào sử dụng mới có những kết luận về giải pháp mới này.

Nhận ngày 8-2-1987

TÀI LIỆU THAM KHẢO

1. **Đào Hữu Chí, Thiều Văn Công**, «Về việc cài đặt tiếng Việt trên các máy vi tính» Báo cáo trong hội thảo «Các hệ xử lý văn bản tiếng Việt trên máy vi tính» lần thứ nhất 19-20/1/1987, trang 19-28.

2. «Cài đặt tiếng Việt vào các hệ máy tính», Trường ĐHBK T.P.Hồ Chí Minh. Tóm tắt báo cáo Hội thảo CHXLVBTVTMVT, trang 5.

3. **Thái Lê Thăng, Nguyễn Doãn Phước, Nguyễn Minh Tuấn**, «Về một phương pháp xây dựng hệ soạn thảo tiếng Việt». Tóm tắt báo cáo tại hội thảo CHXLVBTVTNVT, trang 4.

4. **N. T. Nhàn**, «Kiến trúc chữ Việt trên các máy tính điện tử». Báo cáo trong hội thảo CHXLVBTVTMVT, trang 33—45.

5. Hệ điều hành cho máy vi tính IBM—PC. Avanced MS—DOS 12/1986

ABSTRACT

A PROJECT FOR THE INSTALLATION OF THE VIETNAMESE LANGUAGE ON MICRO COMPUTERS

In this article we will discuss two main project for the installation of the vietnamese language on 16 bit micro computers:

— Storing in RAM in a manner which is as similiar as the display on the screen (using codes in the second half of the ASCII table).

— Stosing in RAM in a manner which is different than the display on the screen (using 10 free codes in the ASCII table).

The first project has been implemented successfultly. However, it has also few essential defects in which the most important is the unwarranty of the vietnamese lexical sorting order. That is why the second project should be preferable. Several suggestions to the implementation of the second project are given. Also given in the annex A the VNCII (Vietnamese Code for Information Interchange) table.

PHẦN PHỤ LỤC A:

BẢNG MÃ VNCII (VIETNAMESE CODE FOR INFORMATION INTERCHANGE)

Mã HEX	Mã DEC	Ký tự điều khiển	Mã HEX	Mã DEC	Ký tự	Mã HEX	Mã DEC	Ký tự	Mã HEX	Mã DEC	Ký tự
000	000	NUL	020	032	(space)	040	064		060	096	— dấu nối từ
001	001	SOH	021	033	!	041	065	A	061	097	a
002	002	STX	022	034	!!	042	066	B	062	098	b
003	003	ETX	023	035	#	043	067	C	063	099	c
004	004	EOT	024	036	€	044	068	D	064	100	d
005	005	ÊNQ	025	037	%	045	069	E	065	101	e
006	006	ACK	026	038	&	046	070	E	066	102	f
007	007	BEL	027	039		047	071	G	067	103	g
008	008	BS	028	040	(	048	072	H	068	104	h
009	009	HT	029	041	)	049	073	I	069	105	i
00A	010	LF	02A	042	×	04A	074	J	06A	106	j
00B	011	VT	02B	043	+	04B	075	K	06B	107	k
00C	012	FF	02C	044	,	04C	076	L	06C	108	l
00D	013	CR	02D	045	−	04D	077	M	06D	109	m
00E	014	SO	02E	046		04E	078	N	06E	110	n
00F	015	SI	02F	047	/	04F	079	O	06F	111	o
010	016	DLE	030	048	0	050	080	P	070	112	p
011	017	DC1	031	049	1	051	081	Q	071	113	q
012	018	DC2	032	050	2	052	082	R	072	114	r
013	019	DC3	033	051	3	053	083	S	073	115	s
014	020	DC4	034	052	4	054	084	T	074	116	t
015	021	NAK	035	053	5	055	085	U	075	117	u
016	022	SVN	036	051	6	056	086	V	076	118	v
017	023	ETB	037	055	7	057	087	W	077	119	w
018	024	CAN	038	056	8	058	088	X	078	120	x
019	025	EM	039	057	9	059	089	Y	079	121	y
01A	026	SUB	03A	058	:	05A	090	Z	07A	122	z
01B	027	ESC	03B	059	;	05B	091	` dấu huyền	07B	123	~ cho chữ đ
01C	028	FS	03C	060	<	05C	092	´ dấu sắc	07C	124	˘ cho chữ ă
01D	029	GS	03D	061	=	05D	093	? dấu hỏi	07D	125	^ cho chữ ô, â
01E	030	RS	03E	062	>	05E	094	~ dấu ngã	07E	126	cho chữ ư, ơ
01F	031	US	03F	063	?	05F	095	. dấu nặng	07F	127	