Tạp chí Tin học và Điều khiển học, T. 19, S. 2 (2003), 147–154

XÂY DỰNG PHƯƠNG PHÁP MÃ HÓA ĐỂ SẮP XẾP DỮ LIỆU CHỮ VIỆT

CAO ĐÌNH THI

Khoa Tin học Kinh tế, Đại học Kinh tế Quốc dân

Abstract. In this paper, we construct some coding methods applied in the arrangement of the data in vietnamese.

Tóm tắt. Trong bài báo này chúng tôi sẽ xây dựng một số phương pháp mã hóa ứng dụng trong việc sắp xếp dữ liệu chữ Việt.

1. MỞ ĐẦU

Vấn đề xây dựng các phương pháp để chuẩn hóa và sắp xếp dữ liệu chữ Việt đã được đặt ra từ lâu. Ngay từ khi máy tính điện tử mới du nhập vào nước ta những người làm công tác giảng dạy và nghiên cứu tin học Việt nam ở trong nước cũng như ở nước ngoài đã nghĩ ngay đến việc phải xây dựng một bộ mã cho tiếng Việt và các chương trình phần mềm phục vụ cho việc soạn thảo, lưu trữ và xử lý dữ liệu chữ Việt trên máy tính. Đến nay đã có một số bộ mã chữ Việt được sử dụng nhiều như ABC, VIETWARE, VNI,... Bất kỳ bộ mã nào cũng đều có những ưu điểm riêng, nhược điểm riêng nhưng nói chung chúng đã đáp ứng được yêu cầu soạn thảo, lưu trữ và in ấn chữ Việt qua máy tính. Hiện nay do việc dùng các bộ mã không đồng nhất ở trong nước cũng như ở nước ngoài (ở miền Bắc thường dùng bộ mã chuẩn TCVN 5712-93 với bộ cài đặt có tên là ABC mà người ta hay gọi tắt là bộ ABC, ở miền Nam hay sử dụng bộ VIETWARE, ở nước ngoài hay dùng bộ VNI) nên việc trao đổi thông tin giữa hai miền và ra nước ngoài gặp rất nhiều khó khăn. Thông thường, người ta phải sử dụng các chương trình chuyển đổi từ bộ mã này sang bộ mã kia rất phiền phức. Để thống nhất việc sử dụng bộ mã chữ Việt trong soạn thảo, lưu trữ, xử lý dữ liệu và văn bản chữ Việt, ngay từ năm 1993 Bộ Khoa học, Công nghệ và Môi trường đã ra quyết định số 2236/QĐ/PTCN qui định dùng bộ ABC trong tất cả các cơ quan Đảng và Nhà nước. Cũng như các bộ mã khác, bên cạnh những ưu việt như font chữ đẹp, dễ soạn thảo, bộ mã ABC vẫn còn một số hạn chế nhất định mà chúng ta cần khắc phục. Khó khăn chính đối với những người xây dựng bộ mã chữ Việt nói chung, bộ mã ABC nói riêng, là ở chỗ ngoài việc sử dụng các chữ cái của hệ La tinh ta còn phải tạo thêm phụ âm Đ, đ, các nguyên âm như Ă, ă, Â, â, Ê, ê, Ô, ô, Ơ, ơ, Ư, ư, và tổ hợp của các nguyên âm này với các dấu thanh (huyền, hỏi, ngã, sắc, nặng). Người ta thường gọi phụ âm này và các nguyên âm trên với các dấu thanh là những chữ cái thuần Việt. Chữ Việt được tạo bởi những chữ cái thuần Việt và các tổ hợp (có nghĩa) của chúng với những chữ cái của hệ La tinh. Vì trong bộ mã ASCII không còn đủ vị trí để thiết kế bộ mã chuẩn cho chữ Việt nên chúng ta đã phải lấy vị trí của một số ký tự ít dùng trong chữ Việt để thiết kế các chữ thuần Việt. Bảng mã ASCII của các chữ thuần Việt trong bộ ABC xem trong [1]. Bên cạnh đó, việc tồn tại song song cùng một lúc 2 font chữ, chữ hoa (có tên bắt đầu bằng .Vn và kết thúc bằng chữ H, ví dụ, .VnTimeH, .VnArialH,... và chữ thường tương ứng với tên không có chữ H ở cuối, ví dụ, .VnTime, .VnArial,...) trong bộ mã ABC không những gây bất tiện cho người sử dụng mà còn gây nhiều khó khăn, phiền toái khi soạn thảo, khi phải sắp xếp hoặc in ấn.

Mặc dù còn một số khiếm khuyết nhưng từ khi được quy định sử dụng chính thức trong các cơ quan Đảng, Chính phủ và Nhà nước Việt nam, bộ ABC đã đóng vai trò rất quan trọng trong việc trao đổi thông tin không những giữa các cơ quan quan trọng này mà nó còn được sử dụng rất rộng rãi trong giới Tin học nước ta. Có rất nhiều văn bản, cơ sở dữ liệu đã được soạn thảo, lưu trữ, xử lý bằng các font chữ của bộ mã này. Từ đó xuất hiện nhu cầu rất lớn về việc chuẩn hóa và sắp xếp dữ liệu chữ Việt được soạn thảo bằng các font chữ của bộ ABC. Muốn đáp ứng được nhu cầu này trước hết chúng ta phải xây dựng các phương pháp, sau đó lập các chương trình phục vụ cho mục đích trên.

Trong bài này chúng tôi sẽ xây dựng một số phương pháp mã hóa, một số thuật toán sử dụng để lập các chương trình sắp xếp dữ liệu chữ Việt trong FoxPro for Windows và trong Microsoft Excel. Ở đây khi nói đến văn bản hoặc dữ liệu chữ Việt chúng ta sẽ hiểu các văn bản hoặc dữ liệu đó được soạn thảo bằng các font chữ của bộ ABC. Vấn đề chuẩn hóa văn bản soạn thảo trong Microsoft Word đã được trình bày ở bài báo [2], chuẩn hóa dữ liệu soạn thảo trong FoxPro ở bài báo [1], trong Excel ở [5]. Trong bài này chúng tôi chỉ xét vấn đề sắp xếp dữ liệu trong FoxPro và trong Excel. Sở dĩ chúng ta xét các dữ liệu trong hai chương trình phần mềm này vì hiện nay có khá nhiều dữ liệu chữ Việt đã được soạn thảo và lưu trữ trong FoxPro và trong Excel. Như trong [3] đã chỉ rõ, những phần mềm nghiệp vụ chính của ngành ngân hàng Việt nam đều được viết bằng FoxPro và hơn 60% dữ liệu của ngành ngân hàng nước ta được soạn thảo và lưu trữ trong FoxPro. Các nguồn dữ liệu này đã được cài đặt trên hơn 500 máy chủ. Còn khối lượng dữ liệu chữ Việt được soạn thảo trên Excel cũng rất lớn. Như chúng ta đã biết, khi làm việc với một tệp dữ liệu bất kỳ, chúng ta phải làm việc với các trường (fields) (hay còn gọi là các cột) và các bản ghi (records) (các hàng) của tệp đó. Trong bài này chúng ta sẽ hiểu dữ liệu chữ Việt là dữ liệu mà giá trị của các trường, các bản ghi được soạn thảo bằng các font chữ của bộ ABC. Trên thực tế chúng ta rất hay gặp các trường hợp phải xử lý thông tin trong các tệp dữ liệu có trường họ tên (hoten) lưu giữ họ, đệm, tên của các đối tượng phải quản lý như khách hàng, học sinh, sinh viên,... Việc xử lý thường là chuẩn hóa, sắp xếp, tìm kiếm theo một tiêu chuẩn nào đó. Ví dụ, hàng năm các trường đại học, cao đẳng phải lập danh sách các thí sinh thi vào trường mình. Để tiện cho việc đánh số báo danh và để cho thí sinh dễ dàng tìm xem mình thi ở phòng thi nào chúng ta phải sắp xếp danh sách thí sinh theo tên, theo thứ tự chữ cái của chữ Việt. Việc quản lý các khách hàng, các nhân viên của một doanh nghiệp, một Công ty cũng có những vấn đề tương tự cần xử lý. Hay một ví dụ khác, khi quản lý một kho hàng hoặc một kho vật tư, nhiều khi chúng ta phải sắp xếp, tìm kiếm theo tên hàng hoặc tên vật tư. Nói chung, vấn đề sắp xếp, tìm kiếm trong một trường dữ liệu theo một tiêu chuẩn nào đó là vấn đề chúng ta thường gặp hàng ngày trong cuộc sống, trong xử lý dữ liệu. Nếu trường dữ liệu đó được soạn thảo bằng tiếng Anh, tiếng Pháp hay một ngôn ngữ nào đó chỉ sử dụng các ký tự Latinh thì việc sắp xếp, tìm kiếm không có gì khó khăn vì trong các chương trình phần mềm đã có sẵn các công cụ để làm việc này. Nhưng nếu trường dữ liệu được soạn thảo bằng chữ Việt nói chung và chữ Việt theo các font chữ của bộ ABC nói riêng thì chúng ta không thể sử dụng trực tiếp các công cụ này được. Do đó xuất hiện nhu cầu rất cấp thiết phải xây dựng các chương trình phần mềm để chuẩn hóa và sắp xếp các trường dữ liệu soạn thảo bằng chữ Việt. Để cho việc trình bày được cụ thể, ở đây chúng ta sẽ xét trường dữ liệu là trường họ tên lưu giữ họ, đệm, tên của người Việt nam.

2. SẮP XẾP DỮ LIỆU CHỮ VIỆT TRONG FOXPRO

2.1. Thứ tự trong chữ cái tiếng Việt

Muốn sắp xếp một trường dữ liệu kiểu ký tự trước hết ta phải xác định được thứ tự của

các ký tự tạo lên giá trị của trường đó. Đối với các dữ liệu được soạn thảo bằng các ký tự của hệ Latinh thì thứ tự đó được xác định bằng thứ tự của các chữ cái của hệ này. Nhưng như trên đã trình bày, chữ Việt ngoài việc sử dụng các chữ cái của hệ Latinh chúng ta còn có các chữ cái thuần Việt và thứ tự sắp xếp của chúng lại khác so với hệ Latinh (còn thứ tự các chữ cái của hệ Latinh mà chúng ta sử dụng vẫn theo thứ tự của hệ đó). Dựa vào chỉ dẫn trong [4] chúng tôi xác định thứ tự của các chữ cái trong tiếng Việt như sau:

A,a,à,ả,ã,á,ạ,Ă,ă,ằ,ẳ,ẵ,ắ,ặ,Â,â,ầ,ẩ,ẫ,ấ,ậ,B,b,B,c,D,d,Đ,đ,E,e,è,ẻ,ẽ,é,ẹ,Ê,ê,ề,ể,ễ,ế,ệ,F,f,G,g,
H,h,I,i,ì,ỉ,ĩ,í,ị,J,j,K,k,L,l,M,m,N,n,O,o,ò,ỏ,õ,ó,ọ,Ô,ô,ồ,ổ,ỗ,ố,ộ,Ơ,ơ, ờ,ở,ỡ,ớ,ợ,P,p,Q,q,R,r,S,
s,T,t,U,u,ù,ủ,ũ,ú,ụ,Ư,ư,ừ,ử,ữ,ứ,ự,V,v,W,w,X,x,Y,y,ỳ,ỷ,ỹ,ý,ỵ,Z,z. (1)

Ta tạm gọi thứ tự này là thứ tự gốc. Trong (1) ta có 126 ký tự. Qua thực tiễn sử dụng chúng tôi thấy sắp xếp theo thứ tự này tạo điều kiện rất thuận lợi cho việc tìm kiếm. Cần nhấn mạnh rằng các phương pháp và thuật toán sắp xếp trình bày trong bài này chỉ phụ thuộc vào thứ tự gốc. Nếu muốn sắp xếp theo một trật tự khác ta chỉ cần thay đổi thứ tự của các chữ cái ở đây mà thôi.

Trên thực tế vấn đề xác định đâu là họ, đâu là đệm, đâu là tên của người Việt nam hiện nay vẫn còn là vấn đề mở. Ví dụ, trong "Nguyễn Trần Thanh Hương" có người cho rằng "Nguyễn" là họ, "Hương là tên", "Trần Thanh" là đệm; nhưng có người lại cho rằng "Nguyễn Trần" là họ kép, "Thanh Hương" là tên kép; nhưng trong "Nguyễn Thị Mai Hương" thì người ta chấp nhận "Nguyễn" là họ, "Mai Hương" là tên kép còn "Thị" là đệm. Hiện nay xu hướng sử dụng họ kép và tên kép rất phổ biến ở nước ta. Người ta thường lấy họ của bố đặt trước, họ của mẹ đặt sau tạo thành họ kép. Như đã thành quy luật, muốn tìm kiếm trong một danh sách họ tên người Việt nam người ta thường tìm theo tên chứ không tìm theo họ như người nước ngoài. Mà tên người Việt nam lại viết sau cùng. Hơn nữa, do các công cụ tìm kiếm trong các chương trình phần mềm hiện có lại tìm từ trái qua phải trong một xâu ký tự nên ta không thể sử dụng một cách trực tiếp các công cụ này để tìm kiếm họ tên người Việt được.

Hiện nay vẫn còn tồn tại những tệp dữ liệu mà trong đó người ta lại tách trường họ tên thành hai trường riêng biệt là trường họ đệm và trường tên (người ta coi chữ cuối cùng là tên, còn lại là họ đệm). Danh sách thí sinh thi vào các trường đại học, cao đẳng là một ví dụ. Làm như vậy không những rất bất tiện khi nhập dữ liệu (làm giảm tốc độ nhập) mà khi in ấn lại không đẹp, không đáp ứng đòi hỏi của chuẩn quốc gia. Hơn nữa, trong trường hợp này nếu sắp theo tên sau đó tìm kiếm thì chỉ một việc tìm trong những người có tên là "Hương" thôi cũng rất vất vả vì có rất nhiều người tên là "Hương". Muốn tìm tiếp theo ta lại phải dò theo trường họ đệm, mà một số họ như họ "Nguyễn" ở Việt nam thì lại vô cùng lớn. Như vậy việc tìm kiếm trong trường hợp này là cực kỳ khó khăn. Trong bài này chúng ta sẽ xét trường hợp giá trị của trường họ tên là một xâu ký tự biểu diễn đầy đủ họ tên của người Việt nam, ví dụ, "Nguyễn Trần Thanh Hương".

2.2. Phương pháp mã hóa

Muốn sắp xếp được đúng trước hết chúng ta phải đưa dữ liệu của trường họ tên về dạng chuẩn quốc gia [4], tức là trong xâu ký tự biểu diễn họ tên người Việt mỗi chữ phải cách nhau một ký tự trống (1 dấu cách), các ký tự đầu tiên trong các chữ phải viết hoa. Chương trình chuẩn hóa dữ liệu trong FoxPro được trình bày trong [1]. Sau khi đã chuẩn hóa , để tạo điều kiện thuận lợi cho việc tìm kiếm họ tên người Việt nam chúng ta phải sắp xếp theo chữ cuối cùng trong xâu ký tự biểu diễn họ tên. Ở đây cần nhấn mạnh là chữ cuối cùng chứ không phải ký tự cuối cùng. Ví dụ, chữ "Hương" chứ không phải ký tự "g". Nếu chữ cuối cùng trùng nhau thì sắp theo chữ trước đó. Cứ như vậy sắp đến hết xâu thì thôi. Để thực

hiện được điều đó ta phải đảo ngược xâu ký tự, ví dụ, "Nguyễn Trần Thanh Hương" thành "Hương Thanh Trần Nguyễn". Sau đó sử dụng phương pháp mã hóa và các công cụ sắp xếp trong FoxPro để sắp theo trường đã mã hóa .

Điều thuận lợi trong FoxPro là chúng ta sử dụng được mã của bộ ASCII. Trong bộ ASCII thứ tự sắp xếp của các ký tự được xác định theo mã của nó. Chúng ta sẽ tận dụng khả năng này để mã hóa các chữ cái trong tiếng Việt theo thứ tự (1).

Nội dung chính của phương pháp mã hóa chúng tôi sử dụng ở đây như sau:

- Tương ứng với 126 chữ cái trong (1) chúng ta sẽ tạo ra một dẫy gồm 126 ký tự trong bộ ASCII có mã tăng dần. Không làm mất tính tổng quát, ta có thể lấy các ký tự có mã từ 1 đến 126.

- Thay thế từng ký tự trong xâu ký tự biểu diễn họ tên người Việt nam (đã đảo ngược) bằng ký tự tương ứng trong dẫy vừa tạo ra. Ta sẽ sử dụng hàm Chrtran() trong FoxPro để làm việc này.

Cuối cùng sử dụng công cụ sắp xếp trong FoxPro để sắp xếp trường ký tự mới tạo ra.

Hàm Chrtran() có cú pháp như sau:

Chrtran(< biểu thức ký tự C1 >,< biểu thức ký tự C2 >,<biểu thức ký tự C3>).

Hàm này cho kết quả là một biểu thức ký tự nhận được từ C1 khi thay thế các ký tự có trong C2 bằng ký tự tương ứng trong C3. Ví dụ, Chrtran("abcdefghac", "ace", "xyz") cho kết quả là "xbydzfghxy", (ở đây a thay bằng x, c bằng y, e bằng z).

Từ đó ta có thuật toán sắp xếp trường họ tên người Việt nam theo thứ tự từ dưới lên.

2.3. Thuật toán

• Mở tệp dữ liệu có trường họ tên cần sắp xếp;

• Thêm 2 trường trung gian, ví dụ, htnguoc và mhtnguoc lưu giá trị của trường họ tên đã đảo ngược và mã của trường này khi đã thay thế bằng các ký tự của bộ ASCII.

2.3.1. Đảo ngược xâu ký tự biểu diễn họ tên

1) Khai báo một mảng M, mảng một chiều có 10 phần tử (số chữ tối đa trong họ tên người Việt nam, ta có thể thêm hoặc bớt số phần tử của mảng), bằng lệnh Dimension M(10);

2) Đối với mỗi bản ghi, tính toán số ký tự trống bên trong xâu ký tự biểu diễn họ tên. Hàm occurs(" ", allt(hoten)) giúp ta làm việc này, sktt=occurs(" ", allt(hoten)). Dựa vào số ký tự trống ta có thể xác định được số chữ có trong trường họ tên của bản ghi hiện thời. Ví dụ, trong "Vũ Bão" thì sktt=1, số chữ bằng 2; trong "Nguyễn Trần Thanh Hương" thì sktt=3. Số chữ bằng 4. Tóm lại, số chữ=sktt+1.

3) Gán từng chữ vào từng phần tử của mảng M theo thứ tự tăng dần của các phần tử của mảng. Trong ví dụ trên M(1)="Nguyễn", M(2)="Trần", M(3)="Thanh", M(4)= "Hương".

4) Tạo xâu mới bằng cách ghép từ phần tử cuối cùng của mảng đến phần tử đầu tiên.

```
For i = sktt+1 to 1 step -1
    htng = htng + allt(M(i))
Endfor
```

5) Thay thế tất cả các giá trị của xâu mới này vào trường htnguoc.

2.3.2. Mã hóa theo các ký tự của bộ ASCII

1) Tạo xâu ký tự CV bao gồm 126 ký tự trong (1) theo thứ tự gốc;

2) Tạo xâu ASC bao gồm 126 ký tự của bộ ASCII

```
For i=1 to 126
```

ASC = ASC + chr(i)

Endfor

(Hàm chr(i) cho biết tên của ký tự có mã là i);

3) Thay thế tất cả các giá trị của trường htnguoc bằng các ký tự tương ứng của xâu ASC vào trường mhtnguoc

REPLACE all mhtnguoc WITH chr(htnguoc,CV,ASC);

4) Sắp xếp theo trường mhtnguoc

SORT TO kq ON mhtnguoc

Tệp kq lưu kết quả đã được sắp xếp.

Dựa theo thuật toán này chúng tôi đã lập chương trình chạy với nhiều loại dữ liệu khác nhau. Kết quả cho thấy trường họ tên được sắp xếp từ chữ cuối cùng, các chữ cái theo thứ tự như trong (1). Nếu chữ cuối trùng nhau thì sắp theo chữ trước đó. Cứ thế sắp xếp cho đến hết. Việc sắp xếp theo cách này rất thuận tiện cho việc tìm kiếm theo tên của người Việt nam.

3. SẮP XẾP DỮ LIỆU CHỮ VIỆT TRONG MS EXCEL

Vì trong MS Excel một số ký tự của bộ ASCII không còn phù hợp nên chúng ta không thể sử dụng trực tiếp bộ mã này để mã hóa các ký tự trong (1). Do đó thuật toán và chương trình có khác một chút. Tư tưởng của phương pháp vẫn là bên cạnh dãy các chữ cái của tiếng Việt được sắp theo thứ tự gốc (1) ta phải xây dựng một dãy các ký tự tương ứng mà sự sắp xếp của chúng tuân theo thứ tự tăng dần. Sau đó thực hiện việc mã hóa các chữ cái trong (1) theo dãy mới rồi sử dụng các công cụ sắp xếp của Excel sắp xếp cột đã mã hóa . Trong Excel ta sẽ coi mỗi sheet là một bảng dữ liệu mà trong đó mỗi cột là một trường, các tên trường được ghi ở hàng đầu tiên của bảng, mỗi hàng là một bản ghi. Từ đó muốn sắp xếp theo một trường nào đó ta chỉ cần chỉ rõ tên của cột chứa trường cần sắp xếp và thực hiện khai báo những thông tin cần thiết. Cũng như trong FoxPro, trước khi sắp xếp chúng ta phải đưa giá trị của cột họ tên về dạng chuẩn quốc gia. Chương trình chuẩn hóa dữ liệu trong Excel được trình bày trong [5]. Việc sắp xếp ở đây chúng ta cũng sẽ sắp xếp từ chữ cuối cùng như trong FoxPro.

Thuật toán

- Mở tệp có bảng dữ liệu với trường họ tên cần sắp xếp;
- Xác định vùng dữ liệu cần sắp xếp và tên cột chứa trường họ tên;
- Đối với mỗi bản ghi xác định số ký tự trống bên trong xâu ký tự biểu diễn họ tên người Việt nam, từ đó tính được số chữ thực tế của bản ghi đó;

a) Đảo ngược xâu ký tự biểu diễn họ tên

1) Tạo một mảng, ví dụ C(10) để lưu các chữ trong xâu ký tự biểu diễn họ tên. Nếu số ký tự trống trong xâu là n1 thì số chữ sẽ là n1+1. Giả sử xâu ký tự là "Vũ Thanh Bình" thì số ký tự trống là 2, số chữ là 3 và C(1) = "Vũ", C(2) = "Thanh", C(3) = "Bình".

2) Xâu ký tự mới Chng được tạo bằng cách ghép các phần tử của mảng C theo chiều từ dưới lên.

For i = n1+1 to 1 step -1

Chng = Chng + Trim(C(i))

Next i

Giả sử xâu ký tự ở cột họ tên là "Vũ Thanh Bình" thì xâu ký tự Chng là "BìnhThanhVũ".

3) Chọn cột để lưu các xâu ký tự Chng;

b) Mã hóa

Trong phần này ta sẽ tạo một bộ mã mới sao cho mỗi chữ cái trong (1) tương ứng với một xâu gồm 2 ký tự được sắp theo thứ tự tăng dần.

1) Tạo xâu ký tự gốc bao gồm 126 ký tự theo thứ tự như (1);

2) Tạo xâu ký tự mới gồm 252 ký tự: a0a1...a9b0b1...b9c0...c9.....m5;

Từ 1) và 2) ta có Bảng 1 biểu diễn sự tương ứng giữa các chữ cái tiếng Việt theo thứ tự (1) và mã của chúng.

3) Mã hóa từng hàng của cột lưu giữ các chuỗi Chng bằng cách tìm kiếm, và thay thế mỗi ký tự trong xâu ký tự gốc bằng 2 ký tự tương ứng. Ví dụ, "BìnhThanhVũ" theo bảng 1 sẽ mã hóa thành "c1f1g5e8j5e8a1g5e8l1k1". Ghi giá trị mã hóa này vào cột bên cạnh của vùng dữ liệu đã chọn.

4) Sử dụng công cụ sắp xếp của Excel để sắp xếp cột đã mã hóa .

Bảng 1. Tương ứng giữa chữ cái tiếng Việt với mã

A	a0	Â	b4	đ	c8	ệ	e2	J	f6	õ	h0	ỡ	i4	u	j8	v	l2
a	a1	â	b5	E	c9	F	e3	j	f7	ó	h1	ớ	i5	ù	j9	W	l3
à	a2	ầ	b6	e	d0	f	e4	K	f8	ọ	h2	ợ	i6	ủ	k0	w	l4
ả	a3	ẩ	b7	è	d1	G	e5	k	f9	Ô	h3	P	i7	ũ	k1	X	l5
ã	a4	ẫ	b8	ẻ	d2	g	e6	L	g0	ô	h4	p	i8	ú	k2	x	l6
á	a5	ấ	b9	ẽ	d3	H	e7	l	g1	ồ	h5	Q	i9	ụ	k3	Y	l7
ạ	a6	ậ	c0	é	d4	h	e8	M	g2	ổ	h6	q	j0	Ư	k4	y	l8
Ă	a7	B	c1	ẹ	d5	I	e9	m	g3	ỗ	h7	R	j1	ư	k5	ỳ	l9
ă	a8	b	c2	Ê	d6	i	f0	N	g4	ố	h8	r	j2	ừ	k6	ỷ	m0
ằ	a9	C	c3	ê	d7	ì	f1	n	g5	ộ	h9	S	j3	ử	k7	ỹ	m1
ẳ	b0	c	c4	ề	d8	ỉ	f2	O	g6	Ơ	i0	s	j4	ữ	k8	ý	m2
ẵ	b1	D	c5	ể	d9	ĩ	f3	o	g7	ơ	i1	T	j5	ứ	k9	y	m3
ắ	b2	d	c6	ễ	e0	í	f4	ò	g8	ờ	i2	t	j6	ự	l0	Z	m4
ặ	b3	Đ	c7	ế	e1	ị	f5	ỏ	g9	ở	i3	U	j7	V	l1	z	m5

Dựa vào thuật toán này chúng tôi đã lập chương trình bằng ngôn ngữ VBA (Visual Basic for Application) rất tiện lợi cho việc tạo một Macro để sử dụng cho các bảng dữ liệu khác nhau trong một tệp dữ liệu của Excel. Nếu đã tạo được Macro cho một bảng dữ liệu thì ta cũng dễ dàng ghi vào thư viện chương trình mẫu trong Excel để sử dụng chung cho bất kỳ một tệp dữ liệu nào khác có cột họ tên được soạn thảo bằng các font chữ của bộ ABC sau khi đã chuẩn hóa theo tiêu chuẩn Việt nam.

4. XÂY DỰNG MACRO CHO CÁC TỆP DỮ LIỆU TRONG EXCEL

Dựa theo thuật toán ở Mục 3 ta có thể xây dựng một Macro dùng chung cho các bảng dữ liệu trong Excel. Để làm được điều đó ta có thể thực hiện các bước như sau:

1) Mở một bảng dữ liệu rỗng (chưa có dữ liệu);

2) Tạo một Macro rỗng và gán cho nó một tên nào đó, ví dụ SxCV. Cách tạo một Macro rỗng có thể làm tương tự như trong [2];

3) Vào Macro SxCV (*Tools, Macro,* chọn Macro SxCV, *Edit*) để soạn thảo chương trình bằng VBA. Trong chương trình có các chữ cái tiếng Việt (xâu ký tự từ (1)) nên ta phải chọn font chữ của bộ ABC để soạn thảo chương trình. Để chọn được font chữ này, trong cửa sổ Macro có tên là Visual Basic chọn *Tools, Options, Editor Format;* ở mục Font chọn một font chữ Việt nào đó, ví dụ .VnTime, chọn cỡ chữ ở mục Size. Ta cũng có thể soạn thảo chương trình trong MS Word sau đó copy vào Macro này. Khi copy chương trình từ Word vào Macro cần kiểm tra xem có đủ 126 chữ cái từ (1) không, (thường hay thiếu chữ "ả" và chữ "ơ") nếu không đủ thì phải bổ sung cho đủ, đúng vị trí. Nhấn Ctrl S để ghi lại với một tên nào đó, ví dụ SxChuViet.xls chẳng hạn;

4) Nhấn Alt Q (hoặc vào *Menu File, Close and Return to Microsoft Excel*) để trở về bảng dữ liệu trong Excel;

5) Ra khỏi Excel;

6) Mở lại Excel; mở tệp SxChuViet, nhấn vào nút *Enable Macros;*

7) Gán cho Macro này một tổ hợp phím (Sortcut Key) bằng cách vào *Tools, Macro,* chọn Macro SxCV, *Options.* Gõ từ bàn phím một chữ cái nào đó, ví dụ chữ "T";

8) Trở về bảng dữ liệu SxChuViet; vào *File, Save As;* ghi tệp này vào thư mục *Microsoft Office\Office\Library* với kiểu là *Microsoft Add-In;* (Thông thường thư mục này ở C: Program Files).

9) Ra khỏi Excel;

10) Vào lại Excel, chọn *Tools, Add-In,* đánh dấu vào ô SxChuViet, OK.

Sau khi đã tạo và ghi Macro này vào thư viện chương trình ta có thể sử dụng nó để sắp xếp một bảng dữ liệu bất kỳ có cột họ tên người Việt được soạn thảo bằng các font chữ của bộ ABC bằng cách chỉ cần mở bảng dữ liệu đó và nhấn Ctrl T rồi làm theo chỉ dẫn là ta có thể sắp xếp cột này (các cột khác cũng chuyển theo) từ chữ cuối cùng theo thứ tự (1).

5. KẾT LUẬN

Trên đây chúng tôi đã trình bày các phương pháp mã hóa và xây dựng các thuật toán để sắp xếp dữ liệu chữ Việt được soạn thảo bằng các font chữ của bộ ABC. Như ta đã thấy, thứ tự sắp xếp chỉ phụ thuộc vào thứ tự gốc mà ở đây được biểu diễn bằng (1). Chỉ cần thay thứ tự của các ký tự trong (1) ta sẽ được một cách sắp xếp mới. Điều cơ bản là chúng ta phải chọn một thứ tự nào đó để sau khi sắp xếp ta có thể tìm kiếm được dễ dàng. Đó mới là mục đích chính của việc sắp xếp. Nếu trường (cột) dữ liệu là tên hàng, tên vật tư,... thì trong thuật toán cũng như trong chương trình chúng ta không phải đảo xâu ký tự trong trường (cột) đó mà thực hiện mã hóa rồi sắp xếp luôn. Cần phải nhấn mạnh rằng các kỹ thuật, các thuật toán ở đây không những chỉ dùng cho các font chữ của bộ ABC mà còn có thể cải tiến để sử dụng được cho các bộ mã khác. Sắp tới có thể chúng ta sẽ dùng bộ UNICODE để soạn thảo chữ Việt. Như ta đã thấy, thứ tự của các chữ cái sử dụng trong bộ này không trùng với thứ tự của chữ cái tiếng Việt. Do đó nếu sử dụng bộ UNICODE hay một bộ mã nào khác chúng ta sẽ phải xây dựng các chương trình phần mềm phục vụ cho việc sắp xếp dữ liệu chữ Việt soạn thảo bằng các font chữ của các bộ mã này. Những kỹ thuật trình bày trong bài này chắc chắn sẽ giúp ích cho chúng ta xây dựng các chương trình phần mềm đó. Lưu ý rằng xâu ký tự ASC được tạo bởi Mục 2.3.2.2) có thể bắt đầu từ một ký tự bất kỳ chứ không nhất thiết phải bắt đầu từ ký tự có mã ASCII là 1 nhưng bắt buộc các ký tự trong xâu này phải có mã tăng dần và số ký tự phải đủ 126. Cũng tương tự như vậy xâu ký tự tạo bởi Mục 3.b.2) có thể bắt đầu từ ký tự khác chứ không nhất thiết bắt đầu từ chữ a nhưng bắt buộc phải có đủ 252 ký tự và các cặp 2 ký tự một này phải có thứ tự tăng dần. Dựa theo các thuật toán trên chúng tôi đã xây dựng các chương trình trọn vẹn và đã

thử nghiệm trên nhiều số liệu thực tế. Kết quả chứng tỏ tính hiệu quả rất cao của các thuật toán này. Độc giả muốn sao chép các chương trình này có thể liên hệ với tác giả theo địa chỉ: cdthi@cfvghn.org.vn

TÀI LIỆU THAM KHẢO

[1] Cao Đình Thi, Chuẩn hóa dữ liệu & sắp xếp chữ Việt trong font chữ in hoa, *Tạp chí Tin học Ngân hàng,* (2) (1999) 27–32.

[2] Cao Đình Thi, Chuẩn hóa văn bản chữ Việt soạn thảo trong Word, *Tạp chí Tin học và Điều khiển học,* **17** (2) (2001) 82–86.

[3] Tạ Quang Tiến, Y2K - những vấn đề cần quan tâm, *Tạp chí Tin học Ngân hàng,* (2) (1999) 3–5.

[4] *Từ điển bách khoa Việt nam 1.* Trung tâm biên soạn từ điển bách khoa Việt nam, Hà nội 1995, trang 8.

[5] Vũ Văn Thái, Cao Đình Thi, Chuẩn hóa dữ liệu tiếng Việt trong Excel, *Tạp chí Tin học Ngân hàng,* (5) (1999) 22–25.

Nhận bài ngày 12 - 11 - 2002