

MỘT SỐ VẤN ĐỀ ỨNG DỤNG CÔNG NGHỆ NHẬN BIẾT KÝ TỰ THÔNG MINH TRONG XỬ LÝ SỐ LIỆU TỔNG ĐIỀU TRA DÂN SỐ VÀ NHÀ Ở NĂM 2009

Mai Văn Cầm^(*)

1. Đặt vấn đề

Cuộc Tổng điều tra dân số và nhà ở gần đây nhất của nước ta được thực hiện vào năm 1999. Thông tin cập nhật về dân số là rất cần thiết đối với Chính phủ cho nhiều mục đích khác nhau như lập kế hoạch phát triển kinh tế - xã hội, kế hoạch xây dựng nguồn nhân lực, xây dựng trường học, đường xá và bệnh viện và thực hiện các chương trình phát triển thiết yếu khác.

Liên Hợp Quốc khuyến nghị các quốc gia thành viên nên tiến hành Tổng điều tra dân số với chu kỳ 5 năm hoặc tối đa là 10 năm. Quyết định 141/2002/QĐ-CP ngày 21 tháng 10 năm 2002 của Thủ tướng Chính phủ phê duyệt định hướng phát triển ngành thống kê đến năm 2010 quy định Tổng điều tra dân số và nhà ở tiến hành 10 năm một lần.

Để thực hiện định hướng đó, Tổng cục Thống kê đã và đang tiến hành nhiều công tác chuẩn bị cho Tổng điều tra dân số và nhà ở tiếp theo vào năm 2009.

Việc nâng cao chất lượng và tính kịp thời của số liệu Tổng điều tra là cần thiết để đáp ứng nhu cầu của người dùng tin. Tính kịp thời của thông tin cũng rất quan trọng đối với giao tế cộng đồng. Xã hội sẽ thừa nhận và sử dụng số liệu nếu chúng thịnh hành và sẽ nhận thức nhiều hơn về tầm quan trọng của công tác thống kê. Kết quả là những nỗ

lực của ngành Thống kê sẽ được cải thiện, chất lượng của các sản phẩm thống kê sẽ được củng cố.

2. Quét hình ảnh và nhận biết ký tự thông minh

Nói chung, Tổng điều tra dân số và nhà ở là một hoạt động thu thập số liệu diện rộng bao trùm toàn bộ lãnh thổ của một quốc gia. Đó là một công việc cần nhiều kinh phí, thời gian và sự tham gia của nhiều cán bộ điều tra ở thực địa và một hệ thống xử lý số liệu hiệu quả và khoa học. Mặc dù kết quả sẽ được phát hành càng sớm càng tốt sau điều tra, song thường vẫn có một khoảng trống đáng kể vì công tác nhập và làm sạch thông tin diễn ra trong một thời gian khá dài. Để rút ngắn quá trình này, cần lựa chọn hoặc tăng cường lực lượng cán bộ và thiết bị hoặc áp dụng công nghệ nhập dữ liệu mới, ví dụ như công nghệ quét hình ảnh.



Công nhân đang nhập dữ liệu bằng công nghệ quét

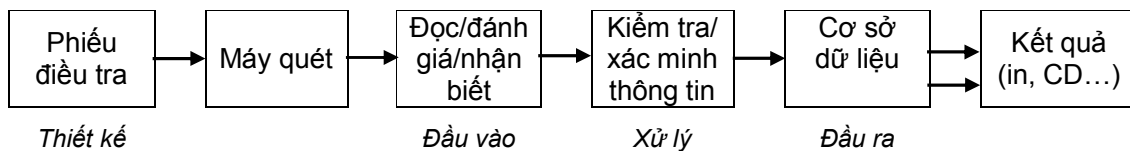
^(*) Phó vụ trưởng Vụ Thống kê Dân số và Lao động

Công nghệ quét hình ảnh là một hệ thống sử dụng để nhập dữ liệu từ phiếu điều tra, máy fax, hoặc internet với tốc độ nhanh và tốn ít nhân lực. Hệ thống này sử dụng máy quét để đọc nhanh phiếu điều tra hoặc tài liệu, và một phần mềm ứng dụng sẽ đọc hoặc định dạng tự động dữ liệu hoặc chữ số ghi trên phiếu/tài liệu, sau đó chuyển thông tin thành file dữ liệu mã ASCII (American Standard Code for Information Interchange) để xử lý tiếp theo. Quét hình ảnh có thể là nhận biết ký tự quang học (OCR – Optical Character Recognition), đọc đánh dấu quang học (OMR – Optical Mark Reader), đọc mã vạch (BCR – Bar Code Reader) hoặc nhận biết ký tự thông minh (ICR -

Intelligent Character Recognition).

Vào thời điểm hiện nay, công nghệ quét hình ảnh được xem là sự lựa chọn tốt cho nhập dữ liệu các cuộc điều tra mẫu quy mô lớn hoặc Tổng điều tra. Công nghệ này giảm thời gian nhập dữ liệu, cần ít nhân lực hơn và giá thành về lâu dài thấp hơn. Trong 4 phương pháp quét hình ảnh nói trên, ICR là lựa chọn phù hợp cho các cuộc điều tra mẫu và Tổng điều tra thống kê, vì phiếu của điều tra thống kê thường được ghi bằng chữ, số và/hoặc khối.

Công nghệ ICR sử dụng phần mềm mạnh cho 4 công đoạn (thiết kế phiếu, đầu vào, xử lý và đầu ra) được trình bày ở hình dưới đây:



3. Ưu điểm của công nghệ nhận biết ký tự thông minh

i. Rút ngắn thời gian nhập dữ liệu với các máy quét tốc độ nhanh. Cần ít nhân lực hơn. Phù hợp đối với các cuộc điều tra mẫu lớn và tổng điều tra, trong đó phiếu điều tra không quá phức tạp, nhưng số lượng phiếu rất lớn. Do nhu cầu tăng cao đối với số liệu kịp thời hơn, công nghệ ICR có thể giúp cho các cơ quan thống kê nhà nước phát hành kết quả điều tra sớm hơn nhiều so với công nghệ truyền thống nhập tin từ bàn phím máy tính.

ii. Công nghệ ICR giảm đáng kể nguy cơ mắc lỗi thủ công trong giai đoạn nhập dữ liệu. Máy quét sẽ đọc phiếu như được ghi trên đó, trong khi nhập thủ công có thể mắc lỗi như đánh thông tin sai hoặc đọc nhầm số liệu. Người ta đã xác định được rằng số

lượng lỗi tăng nhanh khi công việc nhiều. Đối với các cuộc điều tra quy mô lớn, càng ít can thiệp của con người hoặc hoạt động thủ công sẽ giảm được đáng kể sai sót trong giai đoạn nhập số liệu.



Màn hình máy tính khi kiểm tra và xác minh số liệu (verify)

iii. Hệ thống công nghệ ICR cần ít nhân lực hơn so với nhập dữ liệu thủ công.

Máy quét làm việc rất nhanh và đọc thông tin tự động, chỉ đến giai đoạn kiểm tra nhận biết thông tin mới cần người giám sát và hiệu chỉnh.

iv. Do ảnh của phiếu có thể được lưu trữ bằng các file kích thước nhỏ, nên công nghệ này áp dụng rất hiệu quả cho các cuộc điều tra mẫu quy mô lớn hoặc Tổng điều tra. Việc truy cập, tìm kiếm ảnh của phiếu thuận lợi hơn nhiều so với cách tìm thủ công.

v. Chi phí chung của hệ thống công nghệ ICR thấp hơn nhiều so với các công nghệ khác. Chi phí nhập số liệu bằng công nghệ ICR cao ở giai đoạn đầu khi lắp đặt hệ thống, nhưng về lâu dài chỉ cần kinh phí bảo dưỡng và phát triển. Hơn nữa sử dụng ít nhân công và thời gian nhập số liệu ngắn làm giảm khá nhiều chi phí. Kinh nghiệm quốc tế cho thấy rằng, khối lượng công việc càng lớn thì chi phí nhập số liệu bằng công nghệ ICR càng rẻ so với phương pháp nhập thủ công. Điều này càng khẳng định ưu thế của công nghệ ICR đối với các cuộc điều tra quy mô lớn.

vi. Một khi hệ thống được lắp đặt, nó có thể sử dụng cho các cuộc điều tra khác không cần kinh phí đầu tư cho thiết bị.

4. Một số vấn đề cần cân nhắc

i. Mặc dù hệ thống công nghệ ICR có nhiều ưu điểm đặc biệt là đối với các cuộc điều tra quy mô lớn, song có một số vấn đề cần cân nhắc. Ví dụ, việc phân phối và bàn giao phiếu điều tra cần phải được tiến hành một cách cẩn thận, đặc biệt là những địa bàn vùng sâu, xa. Có thể rất khó cho điều tra viên giữ phiếu khô, không ẩm, không nhàu nát. Việc vận chuyển phiếu cần được thực hiện một cách cẩn trọng. Nói cách

khác, phiếu bị gấp mép, ẩm ướt sẽ làm chậm quá trình quét. Vì vậy, phiếu cần được bảo quản trong túi không thấm nước (đối với mỗi địa bàn) và trong hộp các tông (đối với một số địa bàn) để chống ẩm, rách.

ii. Một vấn đề quan trọng là phải kiểm soát được chữ viết tay (tự dạng) của điều tra viên. Viết xấu và ghi không đúng quy định vào phiếu điều tra sẽ gây ra nhiều lỗi khi quét. Đặc biệt là trong Tổng điều tra, một lực lượng lớn điều tra viên sẽ tham gia, thường khó kiểm soát được cách viết của họ. Một trong những nhiệm vụ của giám sát viên là giám sát được tự dạng của điều tra viên.

iii. Chất lượng in của các công ty khác nhau cũng có thể sẽ khác nhau. Thông thường việc in phiếu cho Tổng điều tra sử dụng nhiều công ty khác nhau do số lượng phiếu rất lớn. Khi in số lượng phiếu lớn bằng cách sử dụng nhiều công ty in, cần thu xếp để có được sự kiểm soát chặt chẽ giấy in, chất lượng in, kích thước phiếu in ra, số và ô ghi câu trả lời in trên phiếu. Nếu chất lượng in lô phiếu không đồng đều, như giấy có thể quá mỏng, quá dày, kích thước khác nhau, tốc độ quét sẽ chậm lại do phải thay đổi để giải quyết những vấn đề không thống nhất đó.

iv. Cũng cần phải có kế hoạch sử dụng hệ thống sau khi công việc kết thúc. Có thể xem xét đến khả năng xử lý cho các bài toán khác, cho điều tra giữa kỳ, hoặc làm dịch vụ, như cơ sở dịch vụ cho các đơn vị khác của kinh tế tư nhân hoặc nhà nước. Cũng có khả năng sử dụng hệ thống ICR để xử lý thông tin qua fax hoặc internet. Việc quyết định lắp đặt máy quét tại các tỉnh hoặc vùng có thể được thực hiện trước khi dự án bắt

đầu. Trong trường hợp các cơ sở ở địa phương không có trách nhiệm xử lý số liệu điều tra mẫu hoặc tổng điều tra, thì việc áp dụng công nghệ ICR phi tập trung là không phù hợp. Ngược lại, nếu các cơ sở ở địa phương có hệ thống công nghệ thông tin và có số liệu Tổng điều tra hoặc điều tra mẫu để xử lý, có thể hữu ích nếu lắp đặt máy quét tại đó.

v. Nếu công nghệ ICR được lựa chọn cho Tổng điều tra, hệ thống cần được thử nghiệm với các cuộc điều tra mẫu cũng như với điều tra thử của Tổng điều tra trước khi triển khai chính thức với Tổng điều tra. Tất cả các vấn đề có thể xuất hiện, từ thiết kế phiếu đến xuất bản kết quả, và các khó khăn thường rất khác nhau giữa các quốc gia. Hơn nữa, việc quản lý và giám sát chữ viết tay của điều tra viên, cũng như vận chuyển, bảo quản và bàn giao phiếu điều tra là những công việc không dễ dàng. Các khó khăn nói trên cần được lường trước khi quyết định áp dụng công nghệ đó cho Tổng

điều tra.

5. Kết luận

Hệ thống ICR là công nghệ hiệu quả nhập số liệu đối với các cuộc điều tra mẫu lớn hoặc tổng điều tra. Nếu có thể đảm bảo cung cấp phiếu điều tra với số liệu hợp lý, kịp thời và có chất lượng, hệ thống sẽ cần ít hơn sự can thiệp của con người so với nhập số liệu bằng biện pháp thủ công. Bởi vì Tổng điều tra thu thập số liệu quy mô lớn, hệ thống quét hình ảnh được lắp đặt sẽ có hiệu quả phù hợp xử lý số liệu trong một khoảng thời gian theo yêu cầu. Đối với việc đầu tư lớn như vậy, khả năng sử dụng công nghệ này sau khi Tổng điều tra kết thúc cần được cân nhắc.

Công nghệ ICR sẽ cho phép xử lý số liệu trong thời gian ngắn hơn. Tuy nhiên, một số vấn đề cũng cần phải quan tâm. Chất lượng giấy in phiếu, chất lượng in, phân phối, bảo quản, bàn giao tài liệu cần đến sự quan tâm đúng mức ■