Ứng dụng công nghệ nhận dạng ký tự thông minh (icr) trong xử lý số liệu điều tra thống kê

TS. Thiều Văn Tiến (*)

Từ năm 2006, Trung tâm Tin học thống kê đã nghiên cứu công nghệ nhận dạng ký tự thông minh qua các tài liệu nước ngoài, thông tin trên mạng Internet và qua báo cáo kết quả khảo sát của các đoàn khảo sát tại cơ quan thống kê một số nước trong khu vực. Tuy nhiên, việc nghiên cứu, thử nghiệm công nghệ ICR chỉ thực sự bắt đầu khi Trung tâm được trang bị 01 máy quét Fujitsu fi 5900C và phần mềm FORM 5.2 (gồm 1 module Manager, 1 module Scan, 1 module Transfer, 01 module Interpret 150 và 05 module Verify) của công ty ReadSoft từ sự trợ giúp của Quỹ dân số Liên hợp quốc (UNFPA). Trung tâm Tin học thống kê đã cùng Vụ thống kê Dân số và Lao động thử nghiệm công nghệ nhận dạng ký tự thông minh trong xử lý điều tra biến động dân số, kế hoạch hóa gia đình và nguồn lao động năm 2006; xử lý điều tra biến động dân số, kế hoạch hóa gia đình năm 2007; xử lý điều tra lao động việc làm 1/7/2007; xử lý điều tra thử nghiệm lần 3 của Tổng điều tra dân số và nhà ở. Việc thử nghiệm được thực hiện ở các khâu: thiết kế phiếu, in phiếu, điều tra, quét phiếu điều tra, nhận dạng, kiểm tra lôgic, hiệu chỉnh, chuyển đổi dữ liệu. Phạm vi thử nghiệm được mở rộng dần: lần đầu thử nghiệm trên phiếu điều tra của 01 tỉnh, lần thứ 2 trên phiếu điều tra của 9 tỉnh, lần thứ 3 trên phiếu điều tra của 21 tỉnh. Các lần thử nghiệm sau đều rút kinh nghiệm của các lần thử nghiệm trước đó nên kết quả lần sau thường tốt hơn kết quả của những lần trước đó. Lần thử nghiệm đầu tiên chỉ chứng minh được hệ thống vận hành tốt, còn việc xử lý phiếu thực sự không thành công. Nhiều phiếu không nhận dạng được cả tờ phiếu hoặc từng trường cụ thể. Có nhiều nguyên nhân gây ra tình trạng trên như thiết kế phiếu, chất lượng giấy và bảo quản phiếu, chất lượng in ấn, chất lượng ghi phiếu và cả nguyên nhân chưa có kinh nghiệm trong lập trình. Lần thử nghiệm thứ hai với 2 tỉnh đã có dự kiến và chuẩn bị trước cho công nghệ, việc xử lý bằng công nghệ nhận dạng ICR đạt kết quả khả quan. Bảy tỉnh còn lại kết quả thấp hơn và rất khác nhau giữa các tỉnh. Lần thử nghiệm thứ 3 thực hiện với phiếu điều tra lao động việc làm có kết quả tốt hơn. Dự kiến ban đầu là chọn tất cả những địa bàn viết rõ của cả 64 tỉnh để xử lý bằng công nghệ ICR. Tuy nhiên, do số lượng bản quyền cho việc kiểm tra xác thực dữ liệu hạn chế, máy quét gặp sự cố và thời gian xử lý điều tra ngắn nên chỉ có 21 tỉnh (toàn bộ hoặc một phần) được xử lý với công nghệ ICR. Lần thử nghiệm thứ 4 thực hiện với phiếu điều tra thử nghiệm lần 3 của Tổng điều tra dân số và nhà ở.

Qua thử nghiệm xử lý có thể thấy rằng phương pháp nhập tin bằng công nghệ ICR có **những ưu điểm** như:

1. Chất lượng dữ liệu nhập vào đảm bảo

Khi nhập tin bằng bàn phím, chất lượng dữ liệu nhập phụ thuộc vào những người

(*) Trung tâm Tin học thống kê, Tổng cục Thống kê

nhập tin. Với những cuộc tổng điều tra lớn, thường phải huy động nhiều người tham gia, do vậy dễ xảy ra nhầm lẫn như nhập trùng, nhập sót tập phiếu, nhập thừa hoặc thiếu thông tin đã ghi trên phiếu. Trong khi ứng dụng công nghệ ICR do phiếu chỉ đưa vào máy scanner quét là có thể coi như hết liên quan đến tập phiếu nên những sai sót như trên rất ít xảy ra. Do vậy, khi xử lý bằng công nghệ ICR, những sai sót do người nhập tin gây ra đã bị loại trừ nên chất lượng dữ liệu tốt hơn nhập tin bằng bàn phím.

2. Rút ngắn thời gian xử lý

Thời gian xử lý khi áp dụng công nghệ ICR phụ thuộc vào số lượng, chất lượng phiếu và phụ thuộc vào trang thiết bị dùng trong xử lý. Tuy nhiên với máy quét như máy dùng thử nghiệm tại Trung tâm Tin học trong thời gian qua (tốc độ 100 tờ/phút) thì năng suất lao động bình quân cao hơn rất nhiều so với nhập tin bàn phím. Như vậy, với trang bị hợp lý, thời gian dành cho xử lý điều tra có thể giảm nhiều do máy quét đọc tốc độ cao, tốc độ nhận dạng, tốc độ chuyển đổi kết quả nhận dạng ra file text hoặc cơ sở dữ liệu và tốc độ hiệu đính (verify) cao hơn so với phương pháp nhập tin truyền thống. Do vậy thời gian xử lý thường được rút ngắn so với xử lý nhập tin truyền thống.

3. Có khả năng cung cấp kết quả sớm theo từng phần

Trong nhập tin bằng bàn phím, kết quả tổng hợp chỉ có thể có được khi số liệu đã nhập hoàn chỉnh, mà nhập số liệu lại là khâu kéo dài nhất. Việc cung cấp sớm kết quả chỉ giải quyết được thông qua việc xử lý một cỡ mẫu nào đó và công bố kết quả mẫu trước, nghĩa là việc công bố kết quả sớm bằng cách công bố từng phần bị giới hạn trong phạm vi lựa chọn hẹp. Trong công nghệ quét và nhận dạng, ta có thể chia việc kiểm tra số liệu nhận dạng thành những giai đoạn khác nhau, mỗi giai đoạn chỉ kiểm tra một số trường nhất định. Điều này tạo khả năng cung cấp sớm kết quả theo từng nhóm chỉ tiêu một cách mềm dẻo, hiệu quả.

4. Có thể giảm bớt chi phí kho/mặt bằng chứa phiếu, kệ / giá chứa phiếu và cả nhân lực quản lý, bảo quản phiếu

Các tập phiếu sau khi quét xong đã được lưu giữ đầy đủ dưới dạng ảnh. Vì thế không nhất thiết phải lưu giữ phiếu một thời gian dài sau khi nhập tin. Ngoài khu vực của máy quét, những người kiểm tra dữ liệu chỉ làm việc với máy tính chứ không cần có các phiếu bên cạnh nên yêu cầu mặt bằng ít hơn, giảm bớt các căng thẳng do thiếu mặt bằng làm việc trong các cuộc tổng điều tra. Các tập phiếu không phải bàn giao, chuyển tiếp giữa các khâu khác nhau nên không nhất thiết phải tổ chức các kệ /giá chứa phiếu cho dễ tìm kiếm, không phải tổ chức bảo quản phiếu riêng theo từng công đoạn. Nhân công để bảo quản, sắp xếp, bàn giao phiếu cũng giảm bớt đi rất nhiều.

Tuy nhiên, việc áp dụng công nghệ ICR trong xử lý cũng gặp **khó khăn** như:

1. Kinh phí đầu tư ban đầu lớn

Nếu như việc nhập tin từ bàn phím chỉ yêu cầu các máy PC cho nhập tin có cấu hình tối thiểu thì công nghệ ICR đòi hỏi những máy chủ, máy trạm có bộ nhớ trong và ngoài lớn, tốc độ xử lý cao. Các máy quét cũng phải là những máy công nghiệp có tốc độ cao, chất lượng tốt và nhiều chức năng bổ sung khác và do vậy chi phí cho máy quét lớn. Phần mềm xử lý phiếu tự động là những phần mềm chuyên biệt, thường được bán như một giải pháp chứ không như một phần mềm thông thường. Chính vì vậy, giá phần mềm cao và cũng rất khó lựa chọn đánh giá phần mềm. Do vậy, chi phí đầu tư cho hệ thống thiết bị, phần mềm cao hơn so với nhập tin truyền thống.

2. Chất lượng, hiệu quả của công nghệ phụ thuộc vào chất lượng phiếu điều tra

Để đảm bảo tốc độ và chất lượng số liệu nhập vào, ngay cả việc nhập tin từ bàn phím cũng có những đòi hỏi nhất định đối với phiếu điều tra. Nhưng dù sao, giải pháp nhập tin bằng bàn phím vẫn rất "dễ tính" đối với phiếu điều tra. Xử lý phiếu tự động bằng công nghệ quét và nhận dạng đòi hỏi chất lượng phiếu điều tra cao. Một tập phiếu chất lượng có kém đến đâu thì nó vẫn có thể nhập tin từ bàn phím được cho dù thời gian để nhập tin/sửa phiếu có thể tăng lên vài lần so với những tập phiếu khác. Nhưng một tập phiếu kém có thể rách giấy không phục hồi lại được số liệu, có thể hoàn toàn không nhận dạng được hoặc thời gian kiểm tra/ sửa số có thể tăng lên hàng chục lần.

3. Xây dựng, thiết kế ứng dụng, lập trình tốn nhiều công sức hơn

Thiết kế một ứng dụng cho một mẫu phiếu điều tra cũng tốn nhiều công lao động cao hơn so với giải pháp nhập tin bàn phím. Chính vì thế, giải pháp ICR chỉ có thể có hiệu quả kinh tế đối với những điều tra lớn, hoặc những điều tra định kỳ có mẫu phiếu giữ ổn định trong thời gian dài.

Kết quả nghiên cứu, thử nghiệm công nghệ nhận dạng ký tự thông minh trong thời gian qua đã khẳng định rằng có thể áp dụng công nghệ này trong xử lý điều tra thống kê ở nước ta. Tuy nhiên, việc áp dụng công nghệ ICR trong xử lý điều tra chỉ có thể đạt kết quả mong muốn khi những yêu cầu về mặt thiết bị, phần mềm và yêu cầu về phiếu điều tra được đảm bảo. Có thể nêu một số yêu cầu chính như sau:

1. Về thiết bị và phần mềm

Xử lý điều tra bằng công nghệ ICR phải tuân theo qui trình xử lý. Qui trình này gồm các bước như chuẩn bị phiếu trước khi quét; quét phiếu điều tra bằng máy quét; nhận dạng tự động; kiểm tra, xác thực dữ liệu; chuyển đổi dữ liệu sau khi nhận dạng; hiệu chỉnh tự động; tổng hợp kết quả. Toàn bộ dữ liệu kể từ quét vào máy luân chuyển qua các khâu thực hiện hoàn toàn trong mạng LAN. Trong mạng này, các thiết bị chính bao gồm máy quét (scanner), máy chủ (server), máy trạm (workstasion), thiết bị lưu trữ (backup), máy in và các thiết bị mạng.

Máy quét dùng trong xử lý điều tra thường là máy quét công nghiệp, quét 2 mặt, có tốc độ quét từ 80 đến 160 tờ A4/phút, khay chứa giấy 500 tờ, công xuất tối thiểu /ngày (duty cycle) 30.000 tờ/ngày, có khả năng lựa chọn loại bỏ màu (dropout color). Nếu sử dụng máy quét cấu hình thấp thì rất tốn nhân công thực hiện quét phiếu, thời gian quét kéo dài làm tốn kém thêm các chi phí khác đi kèm. Mặt khác các máy quét tốc độ cao có bộ phận cuốn giấy tốt hơn để có thể làm việc với tốc độ cao, lâu dài và ít kén giấy.

Máy chủ, máy trạm dùng trong xử lý áp dụng công nghệ ICR đòi hỏi phải có cấu hình tương đối cao vì khối lượng thông tin phải xử lý là rất lớn. Máy chủ có số lượng CPU tối thiểu là 2, tốc độ cao, RAM tối thiểu là 8GB, dung lượng ổ cứng tối thiểu 2TB. Máy trạm có RAM tối thiểu 1GB, CPU tốc độ cao, dung lượng ổ cứng tối thiểu 160 GB.

Trong xử lý bằng công nghệ ICR, các máy chủ, máy trạm, máy in đều được lắp thành 1 mạng LAN. Mạng này cần được lắp đặt với các Switch có các cổng tốc độ cao (Gigabit Ethernet) nối tới tất cả các máy chủ, máy trạm thực hiện các công việc quét, nhận dạng, chuyển đổi.

Phần mềm dùng trong công nghệ nhận dạng ký tự thông minh là phần quan trọng nhất trong hệ thống xử lý bằng công nghệ ICR. Các phần mềm được biết đến trên thế giới như phần mềm IFP (Intelligent Form Procesing – Xử lý mẫu phiếu thông minh) của IBM, ABBYY của Nga, Document for FORMS của ReadSoft, TIS (Top Image System) của Israel v.v.

2. Về phiếu điều tra

Khác với xử lý điều tra nhập tin bằng bàn phím, xử lý điều tra bằng công nghệ nhận dạng ký tự thông minh đòi hỏi phiếu điều tra phải đảm bảo yêu cầu ngay từ khi chuẩn bị điều tra (thiết kế phiếu, in phiếu, chuẩn bị bút, lựa chọn và đào tạo điều tra viên) và cả trong khi điều tra (ghi phiếu điều tra, kiểm tra, giám sát quá trình điều tra, đánh ký mã, bảo quản phiếu điều tra). Sau đây là một số yêu cầu chính để đảm bảo chất lượng của phiếu điều tra:

- Thiết kế phiếu điều tra phải đảm bảo yêu cầu quét, nhận dạng. Ví dụ như một trang phiếu phải có các điểm định vị để phân biệt với các trang khác. Độ dài, độ rộng của các ô đánh dấu, ô điền ký tự và khoảng cách giữa các ô phải đảm bảo tuân thủ khuyến cáo của phần mềm nhận dạng.

- Về phiếu điều tra. Yêu cầu chung về giấy in phiếu điều tra là giấy có định lượng bằng hoặc lớn hơn 80 gram/m^2, không quá trơn, đủ độ dai. Không sử dụng giấy chất lượng kém, giấy bẩn. Các loại giấy Carbon cũng không được dùng bởi giấy sẽ nhanh chóng làm hỏng máy quét. Không dùng giấy có mầu nền khác màu trắng, cũng như những loại giấy có gợn sóng, có nếp nhăn.

In phiếu cần lựa chọn nhà in tốt nhất có thể, in tại một nơi và tốt nhất là in một lần (đợt) và cùng một loại máy. Việc in ấn ngoài việc đảm bảo in sắc nét, các phiếu giống nhau tuyệt đối còn phải không được có các vết "bẩn" khác từ máy in, các tờ phiếu xén và đóng chính xác, loại bỏ tất cả những tờ có lỗi.

- Về bút ghi thông tin trên phiếu. Về nguyên lý máy có thể nhận dạng được các ký tự viết bằng bút bi, bút mực, bút chì…theo nguyên tắc viết chân phương, rõ, sắc nét, và không viết tràn ra các ô đã quy định. Tuy nhiên, trong khâu điều tra còn có sai sót cần tẩy, xóa. Nếu mỗi lần tẩy, xóa lại phải chép lại thì rất tốn kém. Có thể dùng bút chì hoặc bút bi để ghi phiếu. Nếu dùng bút chì nên dùng loại bút kim loại 2B (đúng tiêu chuẩn đủ độ mềm để bảo đảm độ rõ, sắc nét khi viết) vì việc sửa các thông tin viết sai sẽ rất đơn giản (tẩy sạch và viết lại). Nếu dùng bút bi thì việc tẩy sạch sẽ thực hiện bằng băng xóa.

- Ghi thông tin vào phiếu phải tuân thủ chặt chẽ những qui định như kiểu ký tự theo mẫu qui định. Chữ viết chân phương, không bay bướm, uốn lượn. Chữ viết gọn trong từng ô, nét chữ phải đều, không bị mờ, không bị thiếu hoặc thừa nét. Khi viết sai, cần thực hiện đúng qui định về sửa chữ ký tự viết sai, v.v.

- Về bảo quản phiếu. Yêu cầu phiếu điều tra phải được giữ sạch, không bị quăn góc, không bị nhàu, không bị ẩm.

Ở các nước trong khu vực đã sử dụng công nghệ ICR trong xử lý điều tra thường bắt đầu áp dụng xử lý tổng điều tra dân số và sau đó mới áp dụng cho các tổng điều tra khác. Và theo qui luật đó, Tổng cục Thống kê đã quyết định áp dụng công nghệ ICR trong xử lý điều tra dân số và nhà ở 1/4/2009. Công tác nghiên cứu, thử nghiệm công nghệ đã được tiến hành trong 2 năm. Kết quả thử nghiệm cho thấy rằng đây là công nghệ hiện đại, có nhiều ưu việt nhưng đòi hỏi phải tuân thủ nghiêm ngặt những yêu cầu nhất định trong tất cả các khâu của quá trình chuẩn bị điều tra, điều tra và thực hiện xử lý. Những yêu cầu này đòi hỏi thay đổi quan trọng trong tư duy, cách thức thực hiện công việc ở tất cả các khâu, các công việc liên quan đến điều tra. Mỗi người tham gia trong khâu nào đó của tổng điều tra nếu tuân thủ đầy đủ những qui định của công việc thì sẽ góp phần vào sự thành công của việc xử lý bằng công nghệ ICR. Đây là công việc khó khăn vì số lượng người tham gia tổng điều tra dân số và nhà ở 1/4/2009 là rất lớn, trình độ và nhận thức khác nhau, nhưng với quyết tâm của toàn ngành, hy vọng tổng điều tra nói chung và xử lý tổng điều tra bằng công nghệ ICR nói riêng sẽ thành công■