Sử dụng phần mềm TIS eFLOW trong xử lý số liệu điều tra thống kê

TS. Thiều Văn Tiến*

Năm 2009, Tổng cục Thống kê đã xử lý thành công Tổng điều tra dân số và nhà ở 2009 bằng việc áp dụng công nghệ nhận dạng ký tự thông minh (ICR) bằng phần mềm TIS eFLOW đã mở ra khả năng áp dụng công nghệ nhận dạng ký tự thông minh trong xử lý số liệu các cuộc điều tra, tổng điều tra khác. Tuy nhiên, để đảm bảo áp dụng công nghệ này, cần phải nắm vững các chức năng của phần mềm TIS eFLOW và cách thức áp dụng nó trong xử lý số liệu điều tra. Bài viết này sẽ trình bày một cách tổng quan phần mềm TIS eFLOW và việc áp dụng phần mềm này trong xử lý số liệu điều tra thống kê.

1. Tổng quan về phần mềm TIS eFLOW

Phần mềm TIS eFLOW (TIS eFLOW Unified Content Platform) là một trong những phần mềm thực hiện thu thập, xử lý dữ liệu tự động được phát triển bởi hãng Top Images Systems (TIS), một trong các nhà cung cấp giải pháp thu thập và xử lý số liệu hàng đầu trên thế giới. TIS đã cung cấp giải pháp xử lý dữ liệu cho Cơ quan thống kê nhiều nước trên thế giới. Phần mềm TIS eFLOW bao gồm các công cụ: TIS eFLOW Integra, TIS eFLOW Freedom và TIS eFLOW Smart. Mỗi công cụ nêu trên là giải pháp thực hiện nhận dạng dữ liệu có cấu trúc, bán cấu trúc và phi cấu trúc. TIS eFLOW Integra nhận dạng dữ liệu có cấu trúc, TIS eFLOW Freedom nhận dạng dữ liệu bán cấu trúc và TIS eFLOW Smart nhận dạng dữ liệu phi cấu trúc.

TIS eFLOW xử lý dữ liệu theo các công đoạn tạo thành luồng xử lý (Workflow). Luồng xử lý dữ liệu cơ bản trong hệ thống TIS eFLOW bao gồm các công đoạn: Quét/lấy ảnh phiếu (Scan/File portal), nhận dạng trang (FormID), định dạng trang thủ công (ManualID), nhận dạng trường (Processing), kiểm tra nhận dạng (Mass Verify), hoàn thiện dữ liệu (Data Verify) và xuất dữ liệu (Export).

- Quét/lấy ảnh phiếu là công đoạn thực hiện việc quét các phiếu trên giấy thành các ảnh hoặc thực hiện việc đẩy các file ảnh có sẵn vào hệ thống xử lý.

- Nhận dạng trang là công đoạn chạy tự động. Nó duyệt qua tất cả các ảnh đã quét vào, so sánh mỗi trang với các ảnh gốc (ảnh phiếu trắng) để xác định mỗi ảnh là trang nào trong bộ phiếu, đồng thời xác định các vùng cần số hoá dữ liệu trên ảnh. Nếu phiếu có một hay nhiều trang không xác định được, nó sẽ bị chuyển tới trạm định dạng thủ công để định dạng trang bằng tay.

- Định dạng trang thủ công là công đoạn thực hiện kéo các khung định vị các trường dữ liệu về đúng chỗ cho các trang không định dạng được ở trạm nhận dạng trang.

- Nhận dạng trường là công đoạn thực hiện tự động việc nhận dạng các trường bằng các công cụ nhận dạng (Engine).

- Kiểm tra nhận dạng là công đoạn thực hiện kiểm tra các chữ cái, chữ số đã nhận dạng được trước đó.

- Hoàn thiện dữ liệu là công đoạn thực hiện nhập bổ sung những chữ số mà công cụ nhận dạng không nhận dạng được và kiểm tra tính hợp lệ, hợp logic của dữ liệu.

- Xuất dữ liệu là trạm chạy tự động thực hiện xuất dữ liệu đã hoàn thành trong hệ thống TIS eFLOW ra ngoài hệ thống.

Ngoài các trạm trên, TIS eFLOW còn có các trạm giám sát hệ thống (Controller), trạm ngoại lệ (Exception), trạm tự động giản đơn (SimpleAuto). Trạm giám sát thực hiện việc theo dõi hoạt động của các trạm trong hệ thống xử lý và giải quyết các lỗi phát sinh như xóa các bộ phiếu không thể xử lý được, chuyển ngược các phiếu về các trạm trước để xử lý lại khi cần thiết, v.v... Trạm ngoại lệ dành cho việc xử lý các trường hợp ngoại lệ. Trạm tự động

** Trung tâm Tin học Thống kê Khu vực I*

giản đơn để thực hiện những chức năng bổ sung theo yêu cầu của từng hệ thống xử lý.

2. Ứng dụng xử lý số liệu điều tra bằng phần mềm TIS eFLOW

TIS eFLOW là phần mềm lõi để nhập liệu bằng công nghệ quét, nhận dạng ký tự thông minh. Tuy nhiên, để thực hiện việc xử lý, cần phải xây dựng ứng dụng xử lý dữ liệu bằng những công cụ có trong phần mềm TIS eFLOW. Công việc xây dựng ứng dụng bao gồm việc tạo ra một ứng dụng mới; thiết lập các công cụ nhận dạng (Visual Engine, OCR's Engine) dùng trong ứng dụng trên cơ sở các công cụ đã được cung cấp và thiết lập luồng xử lý dữ liệu (hay còn gọi là sơ đồ các trạm làm việc).

+) Tạo ra một ứng dụng

Người ta sử dụng môđun Enterprise Manager của TIS eFLOW để tạo ra một ứng dụng mới, đồng thời thiết lập các thuộc tính cho ứng dụng. Các thuộc tính cho ứng dụng mới cần thiết lập bao gồm: chọn máy chủ để chạy ứng dụng; thiết lập các trạm chạy tự động; thiết lập các nhóm, người dùng; thiết lập các thuộc tính cơ sở dữ liệu thống kê (tên cơ sở dữ liệu, mật khẩu, người sử dụng, máy chủ chứa cơ sở dữ liệu và các bảng cần thiết trong cơ sở dữ liệu).

Trong quá trình tạo ứng dụng mới, người ta phải thiết lập các luồng công việc (Flow), lập các mẫu biểu (Form), lập các trang (Page), lập các trường (Field), lập các nhóm trường (field group), lập các bảng (Table) cho ứng dụng; đồng thời cũng lập các luật (Rule) và hàm kiểm tra lôgic (Validation Function).

+) Thiết lập các công cụ nhận dạng(Engine)

Trong mỗi ứng dụng của TIS eFLOW đều phải thiết lập công cụ nhận dạng. Công cụ nhận dạng bao gồm công cụ nhận dạng ký tự quang học (OCR Engine) và công cụ nhận dạng ảo (Virtual Engine).

- Công cụ nhận dạng ký tự quang học dùng để nhận dạng các trường. Đối với mỗi trường cần phải chọn công cụ nhận dạng ký tự quang học có sẵn và thiết lập các thuộc tính của Engine như độ tin cậy nhận dạng, các thuộc tính về chỉnh ảnh, v.v.

- Để cải tiến kết quả nhận dạng người ta thiết lập công cụ nhận dạng ảo. Thay cho việc sử dụng công cụ nhận dạng ký tự quang học đơn, có thể nhóm các công cụ nhận dạng ký tự quang học giống nhau thành một công cụ nhận dạng ảo. Đồng thời phải thiết lập một cơ chế bình chọn để công cụ nhận dạng ảo có khả năng nhận dạng tối ưu nhất.

+) Thiết lập sơ đồ các trạm làm việc trong ứng dụng

Sơ đồ các trạm làm việc trong ứng dụng được tạo ra bằng cách chọn các trạm quét phiếu, trạm nhận fiel ảnh, trạm nhận dạng trang, trạm định dạng trang thủ công, trạm nhận dạng trường, trạm kiểm tra nhận dạng, trạm hoàn thiện dữ liệu, trạm ngoại lệ, trạm xuất dữ liệu và trạm giám sát trên thanh công cụ của TIS eFLOW.

Sau khi chọn xong các trạm chúng ta bắt đầu kết nối các trạm lại với nhau bằng cách kéo từ trạm này sang trạm khác tạo liên kết giữa hai trạm (các liên kết đó gọi là Routing Rules). Tùy theo liên kết giữa các trạm mà mỗi liên kết có các thuộc tính khác nhau. Ngoài việc chọn các thuộc tính có sẵn, người ta có thể viết các câu lệnh điều kiện cho mỗi liên kết.

Mỗi trạm đều có các thuộc tính và các sự kiện (Events) đặc trưng cho trạm đó. Các sự kiện còn được dùng để nhúng đoạn chương trình viết trên các ngôn ngữ lập trình C#, VB vào trong hệ thống xử lý. Việc thiết lập thuộc tính của một số trạm như sau:

- Thiết lập thuộc tính cho trạm quét phiếu bao gồm việc xác định máy quét và thiết lập các thuộc tính của máy quét; thiết lập thuộc tính điều chỉnh ảnh; thiết lập chế độ quét phiếu; thiết lập dấu hiệu kết thúc một lô; v.v...

- Thiết lập trạm nhận biết trang để nhận các ảnh từ trạm quét phiếu hoặc từ trạm nhận file và so sánh với các ảnh mẫu để nhận biết trang. Các lô phù hợp giữa trang phiếu và ảnh mẫu được chuyển đến trạm nhận dạng trường; các lô bị sai sẽ được chuyển đến trạm định dạng trang thủ công để chỉnh sửa.

Thiết lập các trạm kiểm tra dữ liệu (bao gồm trạm kiểm tra nhận dạng, trạm hoàn thiện dữ liệu và trạm ngoại lệ). Các công việc cần thực hiện trong giai đoạn xây dựng ứng dụng là thiết kế giao diện cho trạm kiểm tra dữ liệu; thiết lập cách kiểm

tra dữ liệu, tạo các luật kiểm tra logic để kiểm tra dữ liệu; v.v...

- Thiết lập trạm xuất dữ liệu bằng cách dùng trạm xuất dữ liệu tự động hoặc lập trình can thiệp vào trạm xuất dữ liệu. Nếu dùng trạm xuất dữ liệu tự động cần phải thiết lập nơi xuất dữ liệu ra; nếu xuất dữ liệu ra cơ sở dữ liệu thì phải chỉ ra máy chủ chứa cơ sở dữ liệu và chuỗi thông tin kết nối với cơ sở dữ liệu; thiết lập chế độ làm việc của trạm là trực tuyến hay gián tiếp; v.v.

3. Lập trình can thiệp vào các trạm làm việc trong phần mềm TIS eFLOW

Để hệ thống xử lý bằng phần mềm TIS eFLOW làm việc hiệu quả, TIS eFLOW cho phép lập trình can thiệp vào các trạm làm việc. Đây là điểm mạnh của phần mềm này vì khi xử lý phiếu điều tra có khối lượng lớn, nếu chỉ dùng những môđun sẵn có thì tốn rất nhiều công sức để giải quyết những vấn đề phát sinh trong quá trình xử lý. Sau đây là một số công việc cần lập trình trong TIS eFLOW:

- Lập trình ứng dụng ghi lại thông tin về các trạm làm việc vào cơ sở dữ liệu quản lý để kiểm soát hoạt động của cả hệ thống, ví dụ như số lượng lô phiếu đang quét, người quét tại mỗi trạm quét phiếu; khối lượng làm việc của mỗi công nhân trong các trạm kiểm tra nhận dạng, trạm hoàn thiện dữ liệu, v.v...

- Lập trình ứng dụng xóa phiếu cho trạm định dạng trang thủ công cho những phiếu không thể xử lý được để chuyển đến công đoạn nhập tin bằng bàn phím.

- Lập trình ứng dụng cho trạm kiểm tra nhận dạng và trạm hoàn thiện dữ liệu để sử dụng một số phím trong quá trình làm việc với từng ký tự; lập trình tạo các hàm kiểm tra lôgic phức tạp.

- Lập trình ứng dụng cho trạm xuất dữ liệu để xuất dữ liệu đã nhận dạng và file ảnh đã quét theo địa bàn điều tra của đơn vị hành chính.

4. Quy trình xử lý số liệu điều tra thống kê áp dụng công nghệ ICR của TIS eFLOW

Như đã đề cập ở trên, việc áp dụng công nghệ nhận dạng ký tự thông minh của TIS eFLOW chỉ thay thế việc nhập tin bàn phím. Hệ thống xử lý số liệu điều tra thống kê gồm nhiều công việc như giao nhận, chuẩn bị phiếu, nhập tin, kiểm tra và sửa lôgic, hiệu chỉnh, tổng hợp kết quả. Ở nước ta, chất lượng giấy, in phiếu, ghi phiếu, bảo quản còn hạn chế nên tỷ lệ phiếu không nhận dạng được còn cao. Vì thế không thể chép lại phiếu để quét vì vừa tốn công lại dễ tăng thêm sai sót. Như vậy, hệ thống xử lý số liệu điều tra sẽ phải tích hợp cả hệ thống quét, nhận dạng ký tự thông minh và nhập tin bàn phím. Quy trình xử lý số liệu điều tra thống kê có áp dụng công nghệ quét, nhận dạng ký tự thông minh của TIS eFLOW như sơ đồ trang 11.

Quy trình xử lý trên đây là sự kết hợp cả kinh nghiệm xử lý số liệu TĐT DS&NO 2009 và kết quả nghiên cứu về TIS eFLOW trong thời gian qua. Quy trình này đang được sử dụng trong xử lý số liệu Tổng điều tra nông thôn, nông nghiệp và thủy sản năm 2011.

Tóm lại, TIS eFLOW là phần mềm áp dụng công nghệ quét, nhận dạng ký tự thông minh trong xử lý dữ liệu. Phần mềm này cho phép tự động chuyển thông tin trên phiếu điều tra vào trong máy tính, thay thế cho phương pháp nhập tin truyền thống bằng bàn phím. Do vậy, nó chỉ là một phần trong quy trình xử lý số liệu một điều tra thống kê nói chung, nhưng là phần quan trọng nhất vì thường được dùng trong xử lý các cuộc điều tra, tổng điều tra có khối lượng lớn. Tuy nhiên, việc áp dụng phần mềm này chỉ thực sự hiệu quả nếu quy trình xử lý phù hợp và ứng dụng xử lý được xây dựng tốt làm cho hệ thống xử lý có tốc độ và chất lượng xử lý cao./.

TÀI LIỆU THAM KHẢO

1. eFLOW Visual Designer User Guide, TIS Technologies Inc.

2. eFLOW Programmer's Guide, TIS Technologies Inc.

3. eFLOW Run Time Modules, TIS Technologies Inc.

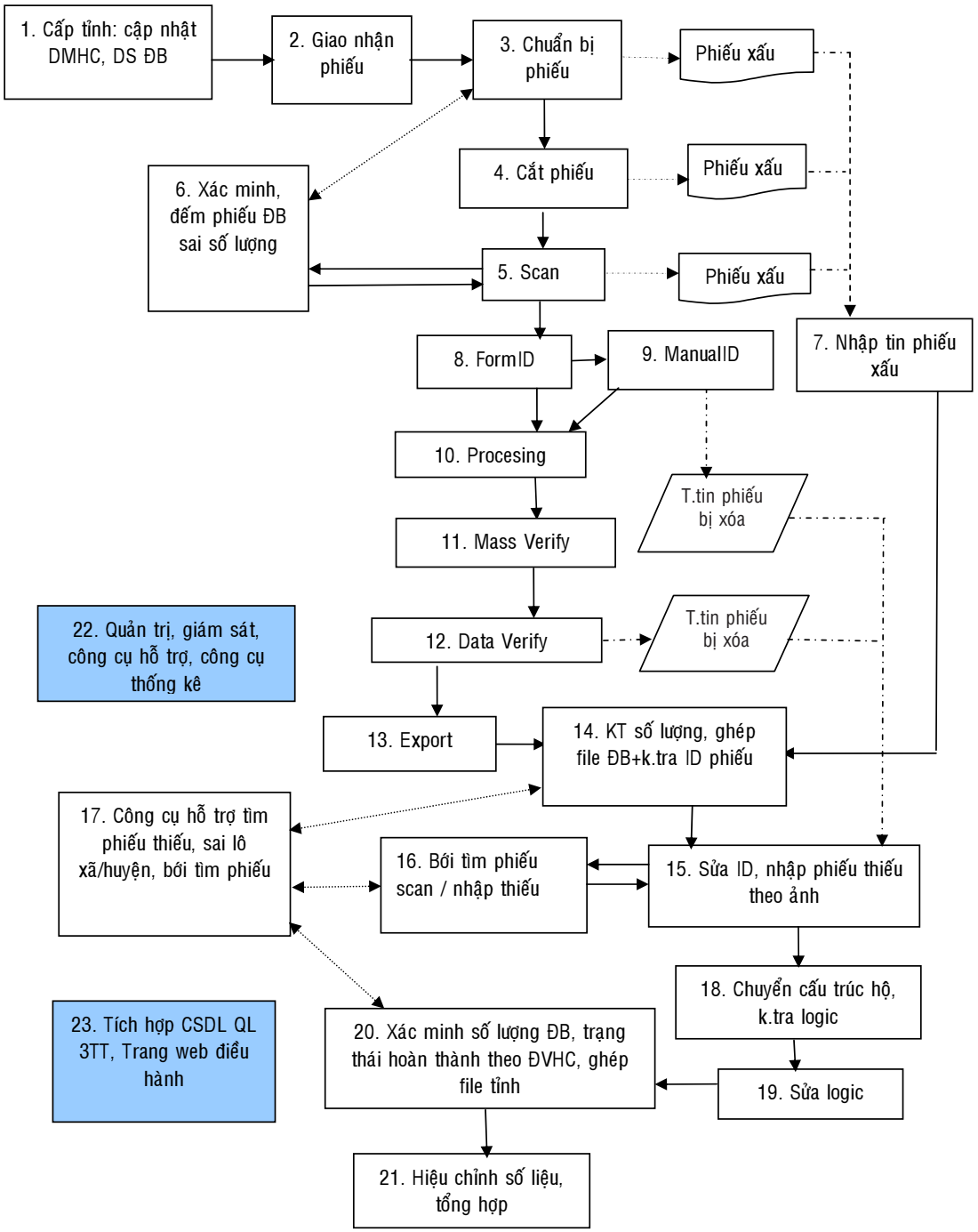
4. eFLOW Installation Guide, TIS Technologies Inc.

5. eFLOW Layout Designer User Guide, TIS Technologies Inc.

6. Data Access Layer Programmer's Guide, TIS Technologies Inc.

7. eFLOW Operator User Guide (Top Image System Singapore)

8. eFLOW Administration Guide (Top Image System Singapore)



Ghi chú:

1. ⟶ Luồng công việc chính, trong đó phiếu hoặc thông tin số hoá được trên phiếu được chuyển từ công đoạn này sang công đoạn khác
2. ⇢ Luồng di chuyển thông tin: mã định danh, ảnh phiếu, ... cho công đoạn xử lý tiếp theo khi cần thiết
3. ⟷ Gửi nhận các thông báo, yêu cầu hỗ trợ, xác minh và thông tin phản hồi