Quá trình ứng dụng phần mềm nhận dạng chữ in tiếng Việt ABBYY ở Trung tâm Thông tin – Thư viện Đại học Quốc gia Hà Nội

Đặt vấn đề

Trên thế giới và Việt Nam có khá nhiều phần mềm quản trị thư viện khác nhau, mỗi phần mềm đều có những tính năng ưu việt phù hợp với điều kiện thực tế của thư viện. Hiện nay một số trung tâm thông tin thư viện đang sử dụng phần mềm nhận dạng chữ tiếng Việt ABBYY là phần mềm được sử dụng rộng rãi trên toàn thế giới. Với những tính năng ưu việt của phần mềm này, một số thư viện ở Việt Nam đã lựa chọn và sử dụng nó để áp dụng cho thư viện mình và Trung tâm thông tin thư viện Đại học Quốc gia Hà Nội là một trong những thư viện đó.

1. Giới thiệu về Trung tâm Thông tin thư viện Đại học Quốc gia Hà Nội

Trung tâm thông tin thư viện Đại học Quốc gia Hà Nội được thành lập theo nghị quyết số 66/TCCP ngày 14/2/1997 của Giám đốc Đại học Quốc gia Hà Nội trên cơ sở hợp nhất của 3 thư viện thành viên. Sau hơn 10 năm xây dựng và phát triển, tới nay Trung tâm được trang bị tương đối đầy đủ nguồn lực thông tin, phương tiện hiện đại đáp ứng nhu cầu của người dùng tin ở nhiều lĩnh vực khác nhau. Trung tâm đã đáp ứng tốt nhu cầu cung cấp tin tri thức cho việc học tập, nghiên cứu khoa học của đội ngũ giảng viên và sinh viên nói riêng, bồi dưỡng nhân tài cho quốc gia nói chung, có nhiệm vụ nghiên cứu, thu thập, xử lí, thông báo và cung cấp tin, tư liệu về khoa học, giáo dục, ngoại ngữ và công nghệ phục vụ cán bộ và sinh viên ĐHQGHN.

2. Tình hình áp dụng phần mềm nhận dạng ABBYY ở Trung tâm thông tin thư viện Đại học Quốc gia Hà Nội

ABBYY có 2 dòng sản phẩm nhận dạng chính: ABBYY Recognition Server và ABBYY FlexiCapture.

Hiện nay Trung tâm thông tin - thư viện Đại học Quốc gia Hà Nội đang sử dụng dòng sản phẩm ABBYY Recognition Server.

Phần mềm ABBYY Recognition Server có thể nhận dạng các tài liệu in của hơn 198 ngôn ngữ với độ chính xác trên 99%, bao gồm cả tiếng Việt; cấu trúc văn bản được giữ nguyên; tốc độ nhận dạng cao, chỉ 2 giây cho 1 trang khổ A4. Sau khi nhận dạng, ABBYY Recognition Server cho phép kết xuất kết quả nhận dạng ra nhiều định dạng file có thể tìm kiếm và biên tập được như MS Word, MS Excel, PDF, PDF/A, HTML, XML.

Trong đó, định dạng PDF/A – một loại định dạng PDF – là định dạng giữ nguyên ảnh quét gốc nên đảm bảo tuyệt đối tính tin cậy của thông tin cho người đọc, đồng thời vẫn cho phép người dùng biên tập và tìm kiếm toàn văn.

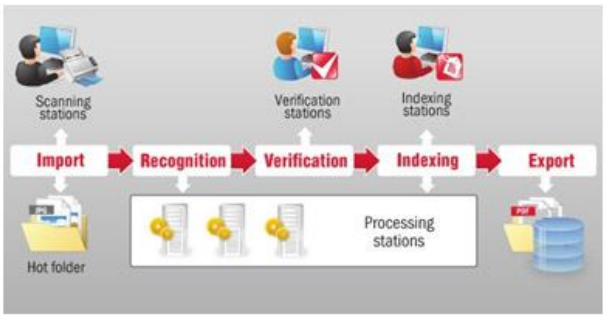
a. Ưu điểm phần mềm ABBYY

- Lưu trữ: Khả năng chuyển đổi một khối lượng lớn tài liệu giấy sang tài liệu số dưới các định dạng có thể tìm kiếm và biên tập được như là MSWord, MS Excel, PDF, PDF/A.
- Nhân viên nhận dạng: Là giải pháp mạnh về công nghệ, hiệu quả về đầu tư cho bài toán nhận dạng văn bản và chuyển đổi dữ liệu của trung tâm. Sau khi cài đặt ở trụ sở chính của trung tâm, mọi nhân viên có thể sử dụng dịch vụ đó tại nhà.
- Tích hợp hệ thống: Nó không chỉ cung cấp giao diện tích hợp dễ sử dụng mà còn là các hàm được đóng gói ở mức độ cao, sẵn sàng cho các chức năng nhận dạng tài liệu hay chuyển đổi sang PDF.

b. Chức năng chính của ABBYY Recognition Server được thực hiện như sau:

+ Nhập ảnh: Trong khâu này Server Manager truy xuất và đọc ảnh từ các nguồn lưu trên File trước: Thư mục chia sẻ trong mạng nội bộ, thư mục FPT, thư mục trong Mailbox trước khi đưa chúng vào hàng đợi để xử lý.
+ Xử lý: File ảnh đến lượt xử lý sẽ được phân luồng xử lý tại trạm xử lý. Nếu hệ thống có nhiều trạm xử lý, Server Manager sẽ phân bổ công việc một cách hợp lý cho các trạm này. Sau khi trạm xử lý nhận dạng xong file ảnh, trả kết quả lại cho Server Manager và tiếp tục nhận file ảnh khác xử lý.
+ Soát lỗi: Nếu chức năng này được thiết lập, những trang cần soát lỗi sẽ được xếp vào hàng đợi sau đó kết quả soát lỗi sẽ được trả về Server Manager.
+ Xuất bản: Sau khi tài liệu được nhận dạng và và chỉnh sửa lỗi, Server Manager sẽ trả tài liệu tới địa chỉ được chỉ định, có thể là một thư mục trên mạng LAN, một thư viện Share point hoặc địa chỉ email.
+ Xử lý lỗi: Các tài liệu có độ nhận dạng tin cậy thấp sẽ được lọc ra và lưu vào một thư mục khác.
+ Khả năng chịu lỗi: ABBYY Recognition Server được thiết kế làm việc hoàn toàn tự động, nó có tính năng đặc biệt để đảm bảo khả năng chịu lỗi và đem đến sự bền bỉ cho hệ thống.

c. Quá trình chuyển đổi tài liệu trong sự công nhận Server có thể chia làm 6 phần hợp lý như sau:



+ Quét/ nhập văn bản: Việc quét Station quét trạm cung cấp chức năng thực thi, chức năng quét và chuẩn bị hình ảnh. ABBYY Recognition Server có thể tự động nhập hình ảnh từ tài nguyên mạng
+ Công nhận: OCR được thực hiện trên một trạm xử lý tự động. Có thể kết nối vài máy tính để quản lý máy chủ như các trạm xử lý, và Server Manager sẽ cân bằng khối lượng công việc trong số các trạm đồng đều.
+ Quản lý chất lượng: Chất lượng quét không thể là hoàn hảo, bị độ phân giải thấp không mong muốn. Trong trường hợp này là rất quan trọng để có một cơ chế bảo đảm chất lượng đáng tin cậy.
+ Tài liệu Separation: ABBYY Recognition Server cung cấp một số tùy chọn tách được xây dựng trong tài liệu: theo trống tờ, tờ mã vạch hoặc in trên trang

đầu tiên của mỗi tài liệu.

+ Thiết loại tài liệu và các thuộc tính: Máy chủ công nhận ABBYY là phần mềm chỉ mục tài liệu mạnh mẽ. Phân loại tài liệu và chỉ mục có thể được thực hiện dễ dàng với diện trực quan chỉ mục Station cho phép các nhà điều hành để chọn một loại tài liệu trong danh sách được xác định trước và điền vào thuộc tính tài liệu.

+ Xuất bản: Sự công nhận máy chủ có thể chuyển đổi hình ảnh thành định dạng tìm kiếm hoặc có thể chỉnh sửa các loại: PDF, PDF/A, RTF, TXT, DOC(X), XLS(X), XML và thành định dạng hình ảnh phổ biến: TIFF, multi-page TIFF và JPEG.

d. Kết quả đạt được và những khó khăn

** Những thuận lợi:* Từ khi áp dụng phần mềm nhận dạng chữ tiếng Việt ABBYY trong hoạt động thông tin - thư viện đã giúp cho quá trình hoạt động nghiệp vụ thông tin - thư viện ngày càng đạt hiệu quả.

Chúng ta hãy tưởng tượng đang có trong tay một số tài liệu giấy như sách, báo, tờ rơi quảng cáo, hợp đồng v.v… máy quét chỉ có thể giúp biến những tài liệu giấy này thành tài liệu dạng ảnh. Với các tài liệu dạng ảnh, chỉ có thể đọc mà không thể biên tập lại chúng trên các hệ soạn thảo điện tử hiện nay; các hệ thống tìm kiếm cũng không thể tìm được các đoạn văn bản trên những tài liệu này. Với việc sử dụng phần mềm nhận dạng chữ in biên soạn lại các tài liệu dạng ảnh sẽ được thực hiện rất đơn giản!

ABBYY Recognition Server là sản phẩm giúp cho Trung tâm ngày càng tăng số lượng thông tin cung cấp cho phép họ quản lý: tài liệu và hình ảnh, tự động, xử lý và lưu trữ thông tin có giá trị. Sử dụng dữ liệu và nắm giữ sản phẩm, giúp cơ quan thông tin - thư viện tiết kiệm tiền bạc và công sức cho nhập dữ liệu. Khi sử dụng phần mềm bố cục của văn bản vẫn được giữ nguyên, giảm thiểu thời gian chỉnh sửa lại văn bản.

** Khó khăn:* Tuy nhiên khi sử dụng phần mềm ABBYY Recognition Server thì Trung tâm cũng gặp phải một số khó khăn như giá thành của sản phẩm tương đối cao, đồng thời đây là phần mềm nước ngoài nên trong quá trình sử dụng gặp nhiều khó khăn khi bảo trì hệ thống.

3. Đề xuất giải pháp nhằm nâng cao hiệu quả ứng dụng phần mềm nhận dạng chữ tiếng việt ABBYY tại Trung tâm Thông tin - Thư viện Đại học Quốc gia Hà Nội

a. Khai thác các tính năng của phần mềm nhận dạng chữ tiếng Việt ABBYY: Tận dụng những tính năng ưu việt, những lợi ích mà phần mềm mang lại cho phép cơ quan thông tin - thư viện quản lý và nắm giữ sản phẩm, tiết kiệm thời gian, tiền bạc, công sức cho việc nhập dữ liệu.

b. Xây dựng và phát triển nguồn tin điện tử: Nguồn thông tin điện tử trở thành nguồn lực quan trọng đối với tất cả các quốc gia trên thế giới, đồng thời đóng vai trò trực tiếp tạo ra của cải vật chất của nền kinh tế xã hội.

c. Nâng cao trình độ đội ngũ cán bộ TT-TV.

• Có kiến thức chuyên môn về nguồn thông tin, bao gồm khả năng đánh giá và chọn lọc các phần mềm và thiết bị thông tin phù hợp.

• Các kỹ năng sử dụng công nghệ thông tin, khai thác thông tin, đặc biệt là biết quản lý, bảo trì dữ liệu khai thác và cung cấp các tài liệu qua mạng.

• Trình độ ngoại ngữ tốt để sử dụng phần mềm có hiệu quả và có khả năng thu thập thông tin, số hóa tài liệu.

d. Mở rộng mối quan hệ hợp tác trao đổi và chia sẻ nguồn lực thông tin: Thư viện cần mở rộng hợp tác với các thư viện và trung tâm thông tin trong và ngoài nước nhằm trao đổi và học tập kinh nghiệm trong quá trình ứng dụng công nghệ thông tin.

Kết luận

Trung tâm thông tin - thư viện Đại học Quốc gia Hà Nội đang ngày càng phát triển và hoàn thiện hơn, đã và đang hướng tới một thư viện có tầm cỡ trong khu vực Đông Nam Á, qua đó góp một phần tích cực vào nhiệm vụ nghiên cứu đào tạo của Đại học Quốc gia.

Hiện nay, trên thị trường Việt Nam có một số phần mềm nhận dạng chữ in (OCR) tiếng Việt và giải pháp số hóa. Tuy nhiên, thực tế sử dụng các phần mềm này cho thấy chúng bộc lộ nhiều hạn chế như không đọc được ảnh màu (chỉ làm việc với ảnh đen trắng), dàn trang của tài liệu sau nhận dạng hay bị vỡ, công suất xử lý thấp, thao tác thủ công, chỉ làm việc với rất ít ngôn ngữ.

Xuất hiện sau các sản phẩm trên thị trường và dù mới chỉ là thế hệ sản phẩm đầu tiên hỗ trợ tiếng Việt nhưng do được thừa hưởng không chỉ những bí quyết công nghệ mà còn cả những tính năng phần mềm chuyên nghiệp của các phiên bản trước, phần mềm nhận dạng phiên bản mới của ABBYY đang được đánh giá là giải pháp nhận dạng tiếng Việt chính xác và toàn diện tại Trung tâm Thông tin - thư viện Đại học Quốc Gia Hà Nội.

_____

Nguyễn Thị Thủy
K53 Khoa TT-TV, trường ĐHKHXH& NV

Nguồn: Tạp chí Thư viện Việt Nam số 5(31) – 2011 (tr.32- 34)