TRÊN ĐƯỜNG TIẾN TỚI PHÂN LOẠI TỰ ĐỘNG

Vũ Văn Sơn
Hội TT-TL KHCN Việt Nam

Bài viết đề cập nhu cầu phân loại tự động trong thời đại nguồn tin điện tử ngày một gia tăng. Làm rõ khái niệm và nguyên lý của phân loại tự động dựa trên kỹ thuật phân loại văn bản, quan hệ giữa phân loại và tìm tin. Nhấn mạnh sự hỗ trợ của khổ mẫu MARC 21 (CF) và các công cụ định chủ đề, như: thesaurus, Khung đề mục chủ đề,... cho phân loại tự động. Giới thiệu những thành tựu bước đầu của phân loại tự động trong lĩnh vực chuyên ngành, những thách đố cần vượt qua để nâng cao hiệu quả, đảm bảo tính chính xác và tính đầy đủ của phân loại và tìm tin.

Trong thế kỷ trước, vai trò của các khung phân loại thư viện đã được nâng cao như là công cụ để định vị tài liệu trên giá, tìm lướt tài liệu qua Mục lục truy cập công cộng trực tuyến (OPAC), và hiện nay để tổ chức và tiếp cận các nguồn tin số hoá trong môi trường kết nối mạng.

Với sự phát triển của Web, nhiều tổ chức trong đó có các viện nghiên cứu, cơ quan thông tin, thư viện đã đưa nội dung thông tin lên mạng. Tuy nhiên, phần lớn các thông tin này lại thiếu các siêu dữ liệu (thông tin thư mục trong thế giới số) làm cho kết quả tìm kiếm kém chính xác và không đầy đủ. Hơn nữa, nhiều khi, các thư viện cố gắng cung cấp siêu dữ liệu, kèm theo nội dung, nhưng các tập hợp siêu dữ liệu này lại không thống nhất ngay cả trong một lĩnh vực cụ thể, thí dụ, do sử dụng những khung phân loại khác nhau để phân loại nội dung; Ngoài ra, không có một hệ thống thư viện số liên kết với nhau để cung cấp những giao diện đồng nhất cho mọi nội dung xuất phát từ những tổ chức hay thư viện khác nhau. Vấn đề đặt ra là phải nghiên cứu phương pháp phân tách siêu dữ liệu và phân loại nội dung đến từ các thư viện khác nhau tạo điều kiện phát hiện và khai thác các nguồn tin được nhanh chóng, trong đó có việc sử dụng và thích nghi một khung phân loại truyền thống. Với việc đầu tư nghiên cứu khâu phân loại của quá trình này, chúng ta có thể tối ưu hoá tỉ lệ tìm tin chính xác và đầy đủ cho một nhóm người dùng cụ thể bằng cách vận dụng các thông số khác nhau của các thuật toán thích nghi và phân loại khác nhau.

Như vậy, việc tăng nhanh các nguồn tin số hoá và nhu cầu phải có các công cụ tìm kiếm hữu hiệu để quản trị thông tin quá tải đã làm cho người ta quan tâm đến nhiệm vụ phân loại tự động, mong muốn làm giảm sức lao động của con người ở một mức độ đáng kể và thậm chí thay thế ở một tỉ lệ khiêm tốn.

Phương pháp tiếp cận chính để tổ chức thông tin là phân loại tư liệu và thông tin thu thập được theo một tập hợp các lớp (môn loại) đã được định trước và tìm được thông

tin tương thích bằng cách lướt tìm một danh sách các lớp đã được sử dụng. Đây là phương pháp truyền thống để phân loại và tìm được tài liệu dựa vào một khung phân loại thư viện. Phương pháp này đã được khơi dậy trong môi trường số hoá cùng với sự thịnh hành của các thư mục chủ đề (subject directory) và thư mục Web (Web directory). Tuy nhiên, một vấn đề có tính chất thách đố với quan điểm tiếp cận này là thiếu một sơ đồ phân loại có đủ uy tín và mang tính chuẩn mực trong thế giới số. Xu hướng chấp nhận một khung phân loại thư viện thịnh hành có cơ sở lý luận và thực tiễn vững vàng, hiện nay tỏ ra có triển vọng và tiềm năng để lấp chỗ trống này và mở ra một hướng nghiên cứu mới đầy hứa hẹn cho việc tổ chức và tìm kiếm nguồn tin số hoá.

Đã có một số công trình nghiên cứu liên quan về tính khả thi của việc sử dụng Hệ thống phân loại của Thư viện Quốc hội Hoa Kỳ (LCC) và Khung phân loại thập phân Dewey làm công cụ cho phân loại tự động nguồn tin số hoá song hành cùng với các dự án xây dựng thư viện số, thí dụ, Dự án Pharos cho Thư viện Alexandria dựa vào LCC; Scorpion của OCLC sử dụng DDC; DISIRE của Liên minh Châu Âu sử dụng Khung phân loại thông tin kỹ thuật (EIC) kết hợp với DDC có độ chính xác từ 55% đến 80% với việc áp dụng thành công phương pháp máy học (ML- Machine learning) vào lĩnh vực tìm tin [2].

Mặc dù phương pháp phân loại thủ công, truyền thống chính xác hơn, song phân loại tự động dựa trên siêu dữ liệu có thể giúp ta xử lý các sưu tập tư liệu khổng lồ với hàng triệu đầu sách.

Kỹ thuật phân loại có hai loại cơ bản: kỹ thuật thống kê và kỹ thuật dựa vào kiến thức. Kỹ thuật thống kê có lợi thế đối với các sưu tập lớn, nhằm nhận dạng và tách ra những từ chỉ mục hữu ích. Kỹ thuật này được sử dụng ngày càng nhiều trong phân loại văn bản, bao gồm các mô hình hồi qui đa biến, mô hình xác xuất Baye, cây quyết định, phương pháp tính các từ lân cận, mạng noron.... Trong khi đó, kỹ thuật dựa vào kiến thức căn cứ vào một cơ sở tri thức (knowledge base) hiện thực, các mạng ngữ nghĩa hoặc khung tình huống (case frame), hệ chuyên gia,... Sáng kiến lưu trữ mở (OAI) cũng có thể là một cấu trúc thu thập siêu dữ liệu nhằm tạo điều kiện dễ dàng cho việc khám phá nội dung ẩn chứa trong các tài liệu lưu trữ được phát tán [3, 4].

Như là một nhánh nghiên cứu tương đối mới, xuất phát từ lĩnh vực tìm tin, Phân loại văn bản (PLVB: Text classification) là nhiệm vụ phân loại theo một tập hợp lớp định trước mà không có sự tham gia của con người. Nhiệm vụ này hoàn toàn giống như một nhánh của biên mục chủ đề trong thư viện truyền thống, nhưng khác biệt ở chỗ không phải do các chuyên gia phân loại thực hiện một cách thủ công. PLVB đã trở nên hấp dẫn hơn bao giờ hết do nhu cầu cấp thiết phải có các công cụ tổ chức một khối lượng thông tin số hoá khổng lồ.

PLVB là đánh dấu văn bản với những lớp tương thích nhất được chọn từ một nhóm lớp dự kiến. Có 3 thành phần cơ bản tham gia vào quá trình PLVB. Thành phần thứ nhất là đối tượng được phân loại, đó chính là các tư liệu dưới dạng văn bản. Thành phần thứ hai là các lớp mục tiêu có liên quan. Thành phần thứ ba là thuật toán ánh xạ đóng vai trò là người phân loại (trong trường hợp này là máy tính đang thực hiện nhiệm vụ phân loại tự động). Thuật toán ánh xạ tiếp nhận một tư liệu như là đầu vào và đưa ra một quyết định có tính nhị nguyên: tư liệu có thể rơi vào

một lớp tương thích (phân loại có kết quả) và cũng có thể không tìm được một lớp thích hợp (không có kết quả). Thuật toán phân loại hoạt động như một hộp đen biến dữ liệu đầu vào (tài liệu chưa phân loại) thành dữ liệu mục tiêu (ký hiệu phân loại) ở đầu ra.

Điều quan trọng là làm thế nào cho một hệ thống máy tính tiếp thu được kiến thức cần thiết để phân loại đúng. Phương pháp phổ biến nhất để giải quyết vấn đề này là kỹ thuật ML. Bản chất học tập đại cương của phương pháp này là máy tiếp thu kiến thức từ kinh nghiệm quá khứ. Ở một mức độ nào đó, phương pháp máy học giống như quá trình học tập của con người. Con người có thể tiếp thu kiến thức nhờ đọc tư liệu, còn máy trong hệ thống học máy tiếp thu kiến thức về một đề tài hay một lớp từ những tư liệu đã được các chuyên gia chuyên ngành chọn lọc trước [2].

Kỹ thuật máy học đã được áp dụng để phân loại các dạng tài liệu và dữ liệu trong lĩnh vực y tế (1996), quốc phòng (1998), luật pháp (2001), thực vật (2002), và tư liệu trên Web (2001),... Phân loại tự động cũng được áp dụng cho các phạm trù tài liệu khác không mang tính chủ đề như các thể loại xã luận, báo cáo, tổng quan, kết quả nghiên cứu và trang chủ (homepage), sàng lọc các thư rác (spam-mails).

Vừa qua, việc sử dụng các khung phân loại truyền thống như: LCC, DDC, UDC, NLM,... đã được mở rộng vào môi trường trực tuyến, tuy nhiên, phân loại tự động còn đang đứng trước những thách đố sau đây:

- Các lớp có khối lượng quá lớn và nói chung hay biến động (có khoảng 100.000 lớp/môn loại khác nhau trong LCC và DDC; Các lớp lại thường xuyên được xem xét lại và chỉnh lý), do đó, việc chuẩn bị dữ liệu cho mỗi lớp và xây dựng một hệ thống PLVB tương ứng với mỗi lớp một cách lôgích gặp khó khăn. Ngoài ra, kinh nghiệm cho thấy không phải toàn bộ các lớp trong khung phân loại đều được sử dụng trong thực tế.

- Các khung phân loại tuy đều lấy chủ đề làm đặc trưng cơ bản cho các lớp, nhưng cấu trúc và hệ thống ký hiệu của các khung này lại hoàn toàn khác nhau.

Để giải quyết tình trạng này, người ta thường chi tiết hoá các lớp tuỳ theo đề tài rộng hay hẹp khi áp dụng PLVB: Đối với các sưu tập chuyên đề thì mới sử dụng đầy đủ các cấp độ chi tiết của lớp thuộc chuyên đề ấy. Tài liệu về các đề tài liên quan hoặc rộng thì có thể sử dụng các lớp bên trên của hệ thứ bậc.

Ngoài ra, khi thử nghiệm phân loại tự động, cũng có những khó khăn nhất định trong việc máy tiếp thu các nguồn tri thức. Điều này có liên quan đến tập hợp dữ liệu đưa vào hệ thống, bao gồm kinh nghiệm và kiến thức nền, qui ước gọi là dữ liệu tập huấn (training data). Việc chọn lọc, bổ sung dữ liệu tập huấn thường tốn nhiều công sức và tiền của và có thể không khả thi trong một số trường hợp vì kiến thức nền là một tập hợp dữ liệu tổng quát, có thể áp dụng cho các chủ đề rộng hơn rất nhiều, trong khi đó kinh nghiệm thường bổ ích và dùng cho các chủ đề hoặc lớp cụ thể.

Như vậy, việc nghiên cứu PLVB phụ thuộc vào dữ liệu tập huấn. Hiện nay, đã xuất hiện một số tập hợp dữ liệu chuẩn mực được sử dụng như mẫu chọn điển hình cho nghiên cứu. Việc phát triển nhiều dữ liệu tập huấn thuộc nhiều chủ đề khác nhau có thể tạo điều kiện nghiên cứu có hiệu quả hơn và

dẫn tới những cải tiến lớn trong PLVB. Do hiếm dữ liệu tập huấn và khó tiếp cận với chúng, việc nghiên cứu phát triển các công cụ tạo ra kiến thức nền có ý nghĩa quan trọng, đó là việc sử dụng định nghĩa và quan hệ của các thuật ngữ trong các bộ từ vựng có kiểm soát và các thesaurus. Thực hiện việc tích hợp và liên kết các công cụ tổ chức tri thức và các nguồn khác nhau có thể cải thiện tình hình một cách đáng kể.

Tạo ra dữ liệu là công đoạn tốn kém nhất trong PLVB, nhưng rất may là hiện nay đã sẵn có một khối lượng lớn thông tin số hoá và nhiều công cụ và kỹ thuật xử lý thông tin trong lĩnh vực tìm tin. Có thể khai thác nguồn tin đã được phân loại bởi các chuyên gia trong các hệ thống thông tin như thư mục Web và CSDL trực tuyến. Một khi hệ thống PLVB được thiết lập, nó sẽ tiếp nhận các tư liệu chưa được phân loại như những sản phẩm đầu vào và sản sinh ra các tư liệu đã phân loại ở đầu ra. Lại có thêm khả năng sử dụng lại các tư liệu đã được phân loại này bổ sung vào tập hợp dữ liệu tập huấn.

Một thuận lợi nữa là ngày càng có nhiều nhà nghiên cứu, chủ yếu trong lĩnh vực tin học và thông tin học, quan tâm đến việc phát triển các công cụ, thuật toán, kỹ thuật và phương pháp phân loại tư liệu dựa trên văn bản. Những loại mô hình phân loại khác nhau đang hỗ trợ các hình thức biểu đạt tri thức khác nhau và chấp nhận những phương pháp học tập khác nhau.

PLVB là một nhánh của nghiên cứu tìm tin, đã sử dụng những kỹ thuật và mô hình tìm tin tiên tiến, tận dụng những thành quả của phương pháp và công cụ học tập dựa trên cơ sở trí tuệ nhân tạo. Kỹ thuật và phương pháp xác xuất đã được chú ý nhiều hơn trong hơn một thập niên vừa qua. Thách đố trước mắt là phải vượt lên trên những mô hình hiện nay xử lý những khía cạnh tương đối đơn giản của ngôn ngữ (từ, ngữ và tên) và những mô hình định chỉ mục dựa vào cơ chế đếm đơn giản như tần số và tỉ lệ đồng xuất hiện (co-occurrence) để tiến tới nắm bắt được những khía cạnh cấu trúc và quan hệ ngữ nghĩa có ý nghĩa quan trọng đối với việc phản ánh đặc trưng của các lớp PLVB về mặt chủ đề và đề tài. Một phương hướng có triển vọng là khai thác các mô hình tích hợp các thuật ngữ có kiểm soát tính nhất quán và quan hệ giữa chúng.

Việc sử dụng các khung phân loại truyền thống cho môi trường số hoá đang được chú ý, đầy hứa hẹn và tiềm năng vì những lý do sau đây:

- Có nhiều khung phân loại đã được sử dụng để tổ chức thông tin;

- Có nhiều công cụ hỗ trợ và kết hợp với khung phân loại như các bộ từ vựng có kiểm soát và đề mục chủ đề đã được phát triển và khả dụng;

- Mô tả thư mục (biểu ghi biên mục) hiện đại đã có sự liên kết với nguồn tin toàn văn

Ở nước ta, đã xuất hiện một số tiền đề thuận lợi cho phân loại tự động: Khung phân loại DDC, một trong những công cụ truyền thống đã được các dự án nước ngoài sử dụng trong phân loại tự động, mới được dịch và phổ biến; Khổ mẫu phân loại MARC 21 (CF), một nhân tố quan trọng để tổ chức CSDL phân loại theo một khung phân loại nào đó đã được nghiên cứu [1]; Một số thư viện và trung tâm thông tin đã biên soạn hoặc sử dụng trực tiếp các công cụ kiểm soát từ vựng như bộ từ khoá, LCSH,...; Các dự án thư viện số tiếp tục phát triển với những tìm tòi nghiên cứu về siêu dữ liệu, siêu văn bản và ngôn ngữ đánh dấu mở rộng XML; Chúng ta có khả năng

tiếp thu kinh nghiệm và thành tựu nghiên cứu của nước ngoài trong lĩnh vực phân loại tự động.

Tuy nhiên, con đường dẫn tới phân loại tự động còn nhiều thách đố, vì phân loại là một quá trình xử lý nội dung, trong đó trí tuệ con người đóng vai trò quan trọng trong việc xác định chủ đề của tài liệu và tạo lập ký hiệu thích hợp. Những công trình nghiên cứu thí điểm ở nước ngoài mới thành công chủ yếu trong phân loại tài liệu số hoá chuyên ngành, thông qua siêu dữ liệu và còn chưa đạt được mức độ chính xác tuyệt đối.

Tài liệu tham khảo

1. *Tài liệu tập huấn MARC 21: Phiên bản 2, có bổ sung và chỉnh lý / Vũ Văn Sơn biên soạn. - H.: TVQGVN, 2006.*
2. *Challenges in automated classification using library classification schemes / Kwan Yi // Proc. of 2006 IFLA Annual Conference, Seoul.*
3. *An Automated classification system and associated digital library services / Kurt Maly, ...*
4. *ACM Computing classification system. - http://www.acm.org/class/1998/*

CƠ SỞ CẦN THIẾT ĐỂ XÂY DỰNG HỆ THỐNG CHỈ SỐ TRONG QUẢN LÝ GIÁO DỤC

TS. Vương Thanh Hương

Viện Chiến lược và Chương trình Giáo dục

Đề cập các cách hiểu khác nhau về chỉ số giáo dục. Trình bày cơ sở khoa học của khung liên kết trong việc xác định các chỉ số trong quản lý giáo dục dựa vào kinh nghiệm, khuyến cáo của các tổ chức quốc tế và một số nước trên thế giới có đối chiếu, so sánh với điều kiện thực tế của Việt Nam.

1. Đặt vấn đề

Chỉ số giáo dục (Education Indicator) có nhiều cách hiểu khác nhau. Claude Sauvageot định nghĩa «*chỉ số giáo dục như một công cụ được xây dựng để phản ánh có ý nghĩa về hệ thống giáo dục quốc dân và còn để báo cáo hệ thống đó tới chính phủ, tới nền giáo dục cộng đồng, nói một cách khác là tới toàn xã hội*», còn David Dean trong bài phát biểu về những chỉ số giáo dục và vai trò của hệ thống thông tin quản lý giáo dục thì cho rằng «*chỉ số giáo dục là những số liệu thống kê được dùng để đánh giá các hoạt động của ngành giáo dục, nhưng không phải số liệu thống kê nào cũng là những chỉ số giáo dục*».

Chỉ số giáo dục là thông tin được xử lý để cho phép nghiên cứu các vấn đề của giáo dục. Một hệ thống chỉ số không chỉ là một tập hợp các số liệu thống kê. Hơn thế, nó nhằm đo lường các thành tố khác nhau của một hệ thống giáo dục và cho thấy sự phối hợp chặt chẽ giữa các thành tố đó để giúp cho hệ thống hoạt động. Hệ thống chỉ số cũng cho biết những thay đổi cơ bản trong hệ thống giáo dục qua những thời kỳ khác nhau. Do vậy, hệ thống chỉ số giáo dục không nên hiểu nhầm là một danh sách các tiêu chí với một loạt các bảng số liệu được xử lý cho một cuốn niên giám thống kê giáo