

VỀ NGÔN NGỮ TÌM TIN TIỀN KẾT HỢP VÀ HẬU KẾT HỢP

TS. Lê Văn Việt

Thư viện Quốc gia Việt Nam

1. Tìm tin và hệ thống tìm tin

Hệ thống tìm tin được hiểu là một tập hợp ngôn ngữ tìm tin (với các quy tắc chuyển từ ngôn ngữ tự nhiên sang ngôn ngữ tìm tin và ngược lại) và tiêu chuẩn phù hợp ý nghĩa giữa các mẫu tìm tài liệu và lệnh tìm. Trong các sách báo chuyên môn người ta cũng thường gọi cả những dụng cụ và máy móc chuyên môn chỉ dùng để tìm tài liệu, căn cứ theo các địa chỉ lưu trữ chúng, là các hệ thống tìm tin.

Hệ thống tìm tin được phân ra thành hệ thống tư liệu và hệ thống dữ kiện. Hệ thống tìm tin tư liệu đáp ứng các yêu cầu thông tin đưa vào hệ bằng cách cung cấp các tài liệu gốc, các bản sao hoặc các địa chỉ bảo quản những tài liệu chứa đựng những thông tin cần thiết. Phân hệ của các hệ thống tìm tin tư liệu (chỉ cung cấp các bản mô tả thư mục về những tài liệu cần tìm) đôi khi được gọi là các hệ thống tìm tin thư mục.

Chức năng của hệ thống tìm tin tư liệu là lấy ra khỏi mảng tìm – tập hợp nào đó các tài liệu, những tài liệu có nội dung phù hợp với nhu cầu tin của cá nhân này hay cá nhân khác. Tuy nhiên, nhu cầu tin được phản ánh trong các yêu cầu tin. Vì thế, có thể nói rằng chức năng của hệ thống tìm tin tư liệu là giải đáp (thỏa mãn) yêu cầu tin (chứ không phải nhu cầu tin) bằng cách rút ra từ tập hợp nào đó của các tài liệu (mảng tìm) những tài liệu phù hợp với yêu cầu.

Việc xác định mức độ phù hợp ý nghĩa giữa nội dung tài liệu với nội dung yêu cầu tin chỉ có thể tiến hành trên cơ sở so sánh hai văn bản - mẫu tìm tài liệu và lệnh tìm.

Hiển nhiên, phương pháp hiệu quả nhất để tìm những tài liệu có chứa những thông tin cần thiết trong một thư viện nào đó là đọc qua từng tài liệu của thư viện này. Nhưng trên thực tế, người cán bộ thư viện không thể nào đọc hết được tất cả những tài liệu mỗi khi có yêu cầu thông tin. Vì vậy, người ta buộc phải sử dụng một

phương pháp khác kém hiệu quả hơn, theo đó việc tìm tin được tiến hành không phải theo các văn bản tài liệu, mà theo những đặc tính ngắn gọn về nội dung hoặc theo những dấu hiệu bên ngoài nhất định của tài liệu, mà ngành thông tin – thư viện gọi là mẫu tìm của tài liệu đó. Như vậy, mỗi tài liệu đều được trang bị một mẫu tìm - đặc trưng cho việc thể hiện ngắn gọn nội dung ý nghĩa cơ bản của tài liệu. Đồng thời với việc đó, yêu cầu thông tin cũng được diễn đạt dưới dạng một đặc trưng ngắn gọn, gọi là lệnh tìm.

Như vậy, mẫu tìm tài liệu - là nội dung ý nghĩa cơ bản của tài liệu, được diễn đạt bằng các thuật ngữ của ngôn ngữ tìm tin, (chứ không phải toàn bộ thông tin chứa đựng trong tài liệu đó); nội dung này phù hợp đơn nghĩa với một tài liệu nào đó và dùng để tìm nó trong tập hợp nhiều tài liệu khác.

Mẫu tìm đơn giản nhất của tài liệu là đầu đề của nó. Xem qua đầu đề cuốn sách hoặc bài báo, trong đa số trường hợp độc giả có thể đánh giá được cuốn sách hoặc bài báo đó có đáng chú ý hay không và có nên đọc kỹ hơn hay không. Bản chú thích và bài tóm tắt tài liệu về thực chất cũng là những mẫu tìm.

Lệnh tìm - là nội dung ý nghĩa của yêu cầu thông tin, được diễn đạt bằng các thuật ngữ tìm tin. Vấn đề về tính phù hợp của sự diễn đạt trong yêu cầu thông tin - một nhu cầu thông tin thật sự của nhà nghiên cứu có một ý nghĩa cực kỳ lớn đối với lý thuyết tìm tin.

Việc tìm tin được tiến hành thông qua việc so sánh đơn giản những mẫu tìm tài liệu với lệnh tìm. Nếu mẫu tìm tài liệu trùng với lệnh tìm ở một mức độ cần thiết và đầy đủ thì người ta coi tài liệu đó phù hợp, hay đáp ứng được yêu cầu thông tin. Cách so sánh đó chỉ đúng khi cả mẫu tìm và lệnh tìm đều được diễn đạt bằng một nghĩa và chỉ một mà thôi.

Như vậy khái niệm “phù hợp” là cơ sở để đánh giá hiệu quả kỹ thuật của các hệ thống tìm tin tư liệu. Song “phù hợp” là khái niệm có tính chất chủ quan. Về thực chất, “phù hợp” là thứ phản ứng tâm lý của người dùng tin đối với từng kết quả cụ thể hoạt động của hệ thống tìm tin. Kinh nghiệm cho thấy, tùy thuộc vào vị trí và thời gian, cùng một người dùng tin đánh giá độ “phù hợp” của tài liệu mà hệ thống cung cấp theo cùng một yêu cầu tin có khác nhau. Mặt khác, những người dùng tin

khác nhau có thể đánh giá sự thích hợp của tài liệu do hệ thống đáp ứng cùng một yêu cầu tin theo các cách khác nhau.

Việc lựa chọn tương quan giữa các chỉ tiêu đầy đủ và chính xác của việc tìm phải do các nhiệm vụ của một hệ thống tìm tin quy định. Tuy nhiên, thật khó để xây dựng một hệ thống tìm tin đảm bảo được sự trùng hợp hoàn toàn (100%) giữa mẫu tìm và lệnh tìm. Trên thực tế, với sự phù hợp từ 25% trở lên của mẫu tìm và lệnh tìm, người ta đã coi cuộc tìm tin đó đạt yêu cầu.

Thiếu sự nhất quán và không chính xác trong việc đánh chỉ số mẫu tìm và lệnh tìm là những nguyên nhân chủ yếu làm cho hệ thống tìm tin tư liệu hoạt động kém hiệu quả. Điều đó lại liên quan đến việc chọn lựa ngôn ngữ tìm tin.

2. Ngôn ngữ tìm tin

Ngôn ngữ là một hệ thống những dấu hiệu có bản chất tự nhiên bất kỳ, thực hiện chức năng nhận thức và giao tiếp trong quá trình hoạt động của con người. Ngôn ngữ có thể là ngôn ngữ tự nhiên, hoặc ngôn ngữ nhân tạo. Ngôn ngữ tự nhiên là ngôn ngữ của con người như một công cụ ý nghĩa và một phương tiện giao tiếp, khác với những ngôn ngữ nhân tạo dùng thay thế cho nó. Điểm đặc biệt cần nhấn mạnh là trong ngôn ngữ tự nhiên không có và không thể có sự phù hợp đơn nghĩa giữa các từ và nghĩa. Trong lời nói thông thường của con người không có câu nào có một ý nghĩa xác định thật chính xác. Khó có thể tìm được hai người mà mỗi từ do họ sử dụng lại có ý nghĩa giống nhau, và ngay trong lời nói của một người thì ý nghĩa của cùng một từ cũng thay đổi trong các thời kỳ khác nhau của cuộc đời. Thực tế cho thấy sử dụng ngôn ngữ tự nhiên trong các hệ thống tìm tin để mô tả đơn nghĩa nội dung cơ bản của các tài liệu và yêu cầu thông tin sẽ gặp rất nhiều khó khăn.

Để thay thế cho các ngôn ngữ tự nhiên trong các hệ thống tìm tin, từ lâu người ta đã sử dụng các ngôn ngữ nhân tạo chuyên dụng, mà thường được gọi là ngôn ngữ tìm tin. Ngôn ngữ tìm tin dùng để diễn đạt nội dung chủ yếu của tài liệu và / hoặc diễn tả chỉ yêu cầu thông tin với mục đích tìm kiếm trong vô số tài liệu những tài liệu đáp ứng yêu cầu thông tin được nêu ra.

Nó phải được lược bỏ tính chất không đơn nhất về ngữ nghĩa của ngôn ngữ thực. Mỗi khái niệm của ngôn ngữ thông tin chỉ được mang một trật tự chặt chẽ các

dấu hiệu tượng trưng, đặc biệt là các đối tượng hoặc loại hình các quan hệ giữa chúng, và ngược lại, mỗi một trật tự các dấu hiệu đặc trưng chỉ được phép có một khái niệm thôi. Như vậy, trong ngôn ngữ tìm tin, các yếu tố từ vựng là các đơn vị đã bị tước bỏ phần lớn ý nghĩa từ vựng, nghĩa là mất tính diễn cảm, hiện tượng đa nghĩa, mỗi liên hệ với ý nghĩa...

Hiện nay, trong lĩnh vực tìm tin tư liệu người ta sử dụng ba loại ngôn ngữ tìm tin cơ bản sau đây:

- Các khung (bảng) phân loại thư viện - thư mục;
- Các khung (bộ) đề mục chủ đề - chữ cái;
- Các ngôn ngữ từ khóa, từ chuẩn.

Các bảng phân loại thư viện - thư mục, bộ đề mục chủ đề - chữ cái (bộ chủ đề) được gọi là ngôn ngữ tìm tin tiền kết hợp; ngôn ngữ từ khóa, từ chuẩn hay còn gọi ngôn ngữ tìm tin hậu kết hợp. Tiêu biểu cho bảng phân loại hiện nay là DDC, UDC, LCC, BBK. Các bộ đề mục chủ đề nổi tiếng nhất hiện nay có LC Subject Headings (Mỹ), RAMEAU (Pháp).

Ngôn ngữ tìm tin tiền kết hợp là những ngôn ngữ mà trong đó các ký hiệu, các tổ hợp từ, câu được diễn đạt theo một mối liên hệ kết hợp (hoặc phụ thuộc) trước khi tiến hành định ký hiệu cho tài liệu.

Việc bổ sung thêm các đơn vị từ vựng mới vào trong các ngôn ngữ tiền kết hợp rất hạn chế: chỉ có thể bổ sung trước khi đánh chỉ số các tài liệu, tức là khi xây dựng ngôn ngữ. Các khả năng xây dựng các đơn vị từ vựng của các ngôn ngữ đó thành các đoạn ngữ đúng (hợp lý) cũng rất hạn chế. Vì thế các khung phân loại thư mục - thư viện và chủ đề - chữ cái không có được “lực ngữ nghĩa” cần thiết để tìm tin một cách hiệu quả theo bất kỳ kiểu yêu cầu nào.

Ngôn ngữ tìm tin hậu kết hợp là ngôn ngữ tìm tin mà trong đó các đơn vị từ vựng sẽ được tập hợp vào mẫu tìm chỉ trong thời gian định ký hiệu cho tài liệu hoặc ngay cả trong quá trình tìm chúng. Điều đó có nghĩa là trước khi định ký hiệu thì các đơn vị từ vựng không hề có mối liên hệ cú pháp nào với nhau. Nghĩa là nó rất giống với bảng chữ cái của ngôn ngữ tự nhiên. Các chữ cái chỉ tạo nên các từ khi viết, còn trước đó thì chúng không có mối liên hệ nào với nhau. Từ 24 chữ cái tiếng Việt ta có

thể kết hợp để tạo nên rất nhiều từ. Từ một số lượng tương đối nhỏ các đơn vị từ vựng của ngôn ngữ hậu kết hợp ta có thể xây dựng được một “mệnh đề” (mẫu tìm) mà trên thực tế có thể diễn đạt bất kỳ một ý nghĩa nào đó. Rõ ràng là các ngôn ngữ hậu kết hợp về mặt ngữ nghĩa sẽ mạnh hơn so với các ngôn ngữ tiền kết hợp.

3. Bản chất định chủ đề và định từ khóa

Đề mục chủ đề

Lần đầu tiên các quy định chủ yếu về việc tạo lập và áp dụng bộ chữ cái chủ đề được nhà thư mục học nổi tiếng người Mỹ C.A.Cutter đưa ra trong cuốn sách của ông *Rules for a dictionary catalogue* xuất bản vào năm 1876. Sau đó được nhà thư mục người Anh, E.J. Coates, phát triển thêm trong tác phẩm của mình *Subject catalogues: heading an structure* được xuất bản năm 1960 tại Luân Đôn.

Năm 1898, Thư viện Quốc hội Mỹ bắt đầu xây dựng các đề mục chủ đề cho kho sách của mình. Từ năm 1909 đến năm 1914 Thư viện Quốc hội Mỹ xuất bản “*Đề mục chủ đề sử dụng cho mục lục kiểu từ điển của Thư viện quốc hội*” mà sau này đổi thành *Đề mục chủ đề của Thư viện Quốc hội*.

Định chủ đề là quá trình xử lý nội dung tài liệu xác định đề tài và các khía cạnh nghiên cứu của nó. Kết quả của việc làm này là tạo ra đề mục chủ đề. Hay nói cách khác, đề mục chủ đề là tên gọi của chủ đề.

Đề mục chủ đề có thể là 1 danh từ, 1 cụm danh từ trình bày 1 cách ngắn gọn chủ đề của tài liệu.

Chủ đề của tài liệu là đề tài chủ yếu chứ không phải toàn bộ nội dung của tác phẩm. Một tác phẩm có thể có nhiều chủ đề và do đó có thể có nhiều đề mục chủ đề khác nhau.

Về mặt cấu trúc, đề mục chủ đề có hai loại: đề mục chủ đề đơn và đề mục chủ đề phức.

Đề mục chủ đề đơn là các đề mục chủ đề bao gồm một thành phần duy nhất là tên của các chủ đề.

Đề mục chủ đề phức gồm tên chủ đề chính và các phụ đề. Các thành tố này được liên kết với nhau theo một cú pháp hay công thức nhất định. Trong tài liệu về

5 nguyên tắc định tiêu đề đề mục, ở phần cú pháp, nguyên tắc 5 có dẫn ra 2 tiêu đề đề mục sau: Tiến hóa – Nghiên cứu -- Anh quốc – Luân Đôn – Lịch sử – Thế kỷ 19 và Hoa Kỳ - - Lịch sử – Nội chiến, 1861 – 1865—Hành quân trên biển -- Tàu ngầm. Theo tài liệu của Vũ Dương Thúy Nga thì các phụ đề được sắp xếp theo tuần tự sau: Phụ đề nội dung, Phụ đề địa lý, Phụ đề thời gian, Phụ đề hình thức¹. Và giữa tên chủ đề chính và các phụ đề cách nhau bằng dấu gạch ngang. Nếu so với công thức này với ví dụ đề mục chủ đề về *Tiến hóa* ở trên thì ta không thể hiểu là tại sao lại có phụ đề nội dung *Lịch sử* giữa phụ đề địa lý (*Anh quốc, Luân Đôn*) với phụ đề thời gian (*thế kỷ 19*). Và giữa các phụ đề có 1 hay 2 gạch ngang? Rồi ở ví dụ đề mục 2 về *Hoa Kỳ* thì chúng ta không biết được rằng đâu là Đề mục chính, đâu là phụ đề nội dung, đâu là phụ đề địa lý, thời gian. Và xét trên nội dung các phụ đề (tôi cứ coi Hoa Kỳ là chủ đề chính đi) thì các phụ đề sau đều là phụ đề nội dung cả, Như vậy có tới 4 (hoặc hơn) phụ đề nội dung sao? Chẳng lẽ chỉ dẫn của IFLA lại rắc rối đến vậy. Và trong đề mục chủ đề do IFLA quy định không có trật tự cố định giữa chủ đề chính và các phụ đề?. Tuy nhiên tôi vẫn thiên về cấu trúc cú pháp trong tài liệu của Vũ Dương Thúy Nga. Nghĩa là giữa tên chủ đề chính và các phụ đề cách nhau bằng dấu gạch ngang; giữa tên chủ đề chính, tên phụ đề với phần bổ sung (giải nghĩa) cho chủ đề chính, phụ đề cách nhau bằng dấu phẩy (,). Đó là cú pháp của ngôn ngữ chủ đề. Có thể chính vì cú pháp này mà ngôn ngữ chủ đề được xếp vào ngôn ngữ tiền kết hợp. Và cũng chính vì cú pháp này mà trước đây có một số chuyên gia cho rằng rất khó có thể tìm tin đề mục chủ đề trên môi trường điện tử.

Theo nguyên tắc 5 định chủ đề của IFLA thì tiêu đề phải diễn tả thật đúng đề tài của tài liệu. Tuy nhiên, khi định chủ đề, người xử lý tài liệu có xu hướng đưa về các chủ đề rộng, chủ đề thống nhất. Chẳng hạn, khi tài liệu nói về giao thông đường bộ thì người xử lý lại đưa nó về tiêu đề giao thông. Và các dạng giao thông khác như giao thông đường sắt, giao thông đường thủy, giao thông hàng không..., nếu theo cách tiếp cận này, chắc cũng sẽ tập trung dưới tiêu đề trên. Điều đó, tạo nên sự phân tán tin rất lớn. Khi có một yêu cầu tin về chuyên đề hẹp, chẳng hạn, về giao thông đường bộ, thì người tìm sẽ tra vào chủ đề *Giao thông* và sẽ nhận được tài liệu về tất cả các dạng giao thông khác...Cũng tương tự như vậy, dưới đề mục chủ đề

¹ Ngoài các phụ đề trên, theo ThS. Nguyễn Minh Hiệp còn có *Tiểu phân mục (hay trong bài viết này gọi là phụ đề) phù động tự do*: Là tiểu phân mục đề tài hay hình thức được dùng chung cho nhiều chủ đề nên được liệt kê một danh sách dùng chung và không xuất hiện trong Khung tiêu đề đề mục chuẩn. Danh sách này được đặt ở phần đầu của một Khung tiêu đề đề mục

thống nhất *Sinh thái học* sẽ tập hợp tài liệu về (dùng cho) các lĩnh vực: Cân bằng tự nhiên; Môi trường; Sinh học và sinh thái; Sinh học môi trường.

Ưu điểm nổi bật nhất của ngôn ngữ chủ đề- là ngôn ngữ chung cho cả người sử dụng và người xử lý. Điều đó làm tăng chất lượng xử lý, tìm tin và làm đơn giản hóa quá trình tiếp cận tới các biểu ghi. Tuy nhiên, cũng có ý kiến cho rằng việc mô tả tài liệu bằng phương pháp chủ đề sẽ có ít đề mục chủ đề hơn nên sự mô tả đó sẽ không thực sự đầy đủ.

Từ khóa, từ chuẩn

Có người (Vikkeri - Mỹ) khẳng định rằng phương pháp định ký hiệu hậu kết hợp đã xuất hiện từ cách đây 3.000 năm tại Sume. Tuy nhiên mãi đến thế kỷ 20 phương pháp này mới bắt đầu ứng dụng rộng rãi. Chẳng hạn, đầu những năm 20, người ta đã dùng phương pháp hậu kết hợp vào ngành địa chất, quản lý nhân sự (trên phích lỗ).

Người đầu tiên sử dụng phương pháp xử lý hậu kết hợp trong công tác thông tin là W.E. Batten. Ông đã thành lập một hệ thống tìm tin về bản mô tả phát minh cho tập đoàn "Imperial Ceramic Industries Lmd". Hệ thống tìm tin của ông vận hành trên phích lỗ. Ông tìm ra phương pháp này khi so sánh 2 bảng tài liệu phát minh về chủ đề A với bảng tài liệu phát minh chủ đề B để chọn ra các phát minh chứa A và B.

Năm 1947, nhà toán học Mooers người Mỹ sáng tạo và đăng ký phát minh hệ thống tìm tin cơ khí hoá, dựa trên việc sử dụng ý tưởng về định ký hiệu hậu kết hợp. Ông đã tạo lập các phích đặc biệt với những lỗ xung quanh mép mà ông gọi là Jatocard cũng như bộ mã (các số ngẫu nhiên) và máy chọn lọc riêng. Ông là người đầu tiên đưa ra các thuật ngữ "Descriptor - từ chuẩn". Theo Mooers thì Descriptor là từ hay tổ hợp từ diễn đạt một khái niệm nào đó rộng hơn hoặc hẹp hơn. Ông cũng đưa ra những yêu cầu đối với từ chuẩn: nghĩa của chúng phải đúng với ngữ cảnh cụ thể, mỗi từ chuẩn phải có lời giải thích phản ánh được ý nghĩa đặc thù của nó trong ngôn ngữ tìm tin đó; việc chọn các từ chuẩn phải đảm bảo hiệu quả cao trong tìm tài liệu theo yêu cầu của một nhóm người dùng cụ thể.

Vào những năm 50 của thế kỷ XX, nhà toán học logic M. Taube và nhân viên thông tin hóa học K. Muarc sáng tạo nên phương pháp mới. Phương pháp này dựa trên quan niệm cho rằng nội dung có ý nghĩa chủ yếu của bất cứ tài liệu và yêu cầu

thông tin nào đều có thể diễn đạt bằng cách chọn ra những thuật ngữ phản ánh nhiều nhất trong tài liệu được xử lý. Những thuật ngữ này gọi là từ khóa. Từ khóa là các từ đủ nghĩa hoặc các cụm từ ổn định biểu thị những khái niệm quan trọng nhất mà tài liệu đề cập tới. Năm 1951, Taube đã đưa ra một hệ thống các từ duy nhất (Unitern) hay còn gọi là từ khóa.

Thực tế có 2 loại từ khóa: từ khóa tự do và từ khóa có kiểm soát.

- Trong định từ khóa tự do, người xử lý thông tin sẽ sử dụng tất cả các thuật ngữ có nghĩa có trong nhan đề, tóm tắt hoặc chú giải, lời nói đầu và cả chính văn nữa làm từ khóa khi mô tả các khái niệm mà tài liệu đề cập đến. Trong ngôn ngữ này không phải kiểm soát các quan hệ giữa các thuật ngữ. Điều đó có nghĩa là cùng một khái niệm có thể được trình bày hoặc xử lý bởi hai hoặc hơn các thuật ngữ khác nhau mà giữa chúng không có các mối liên hệ. Kết quả tất yếu là dẫn tới nhược điểm lớn nhất của ngôn ngữ từ khóa tự do trong tìm tin là giảm độ đầy đủ của kết quả tìm; Mức độ “ồn” trong tìm kiếm lớn.

- Ngôn ngữ từ khóa có kiểm soát: là ngôn ngữ được xây dựng trên cơ sở kiểm soát từ, dùng để quy ước cách dùng từ trong quá trình định từ khóa. Trong các dạng từ khóa kiểm soát có từ chuẩn – là những khái niệm đã được chuẩn hóa, có cấu trúc ngữ nghĩa xác định dùng để mô tả đơn nghĩa nội dung tài liệu và tìm tin trong hệ thống thông tin tư liệu. Về mặt này, từ chuẩn cũng không khác là bao so với đề mục chủ đề. Khác biệt chủ yếu nhất có lẽ ở chỗ các từ khóa thường ngắn hơn đề mục chủ đề (chỉ là từ đơn nghĩa; đôi khi – tổ hợp từ ngắn). Điều quan trọng là giữa các từ này không có mối quan hệ nào. Chúng ta chỉ có thể đoán về chúng. Trừ việc thiếu vắng cú pháp thì giữa ngôn ngữ từ khóa có kiểm soát với ngôn ngữ chủ đề không có một sự khác biệt nào. Chứng minh cho luận điểm này chúng tôi muốn dẫn ra ví dụ được nêu ra trong tài liệu *Những nguyên tắc định tiêu đề đề mục*, Chẳng hạn, có một cuốn sách với tiêu đề *Lịch sử sách và thư viện* sẽ có 2 đề mục nhưng chắc là liên quan đến Việt Nam nên người xử lý đã *chỉ định thêm tiêu đề liên hệ Việt Nam* :

Đề mục: 1. Thư viện – Việt Nam – Lịch sử

2. Sách – Việt Nam – Lịch sử

Trước khi trình bày về định từ khóa cho tài liệu này tôi muốn mở ngoặc một chút: nếu theo cú pháp về tên đề mục chính, phụ đề chúng tôi dẫn ra ở trên thì phụ

đề nội dung *lịch sử* phải được đặt trước phụ đề địa lý *Việt Nam* chứ. Nghĩa là, trật tự đúng của 2 đề mục này, theo chúng tôi, sẽ là:

1. Thư viện – Lịch sử – Việt Nam
2. Sách – Lịch sử – Việt Nam

Trở lại định từ khóa cho tài liệu *Lịch sử sách và thư viện* ta sẽ có 3 từ khóa: *Thư viện, Sách, Lịch sử* và nếu đúng là có liên quan đến (của) Việt Nam thì sẽ thêm từ khóa *Việt Nam*. Như vậy, tài liệu này sẽ có 4 từ khóa: *Thư viện, Sách, Lịch sử, Việt Nam*. Nếu cùng một cuốn sách như vậy nhưng là của Hoa Kỳ hay Anh thì đối với ngôn ngữ chủ đề ta sẽ có đến 4 đề mục chủ đề mới: 1. Thư viện – Lịch sử – Hoa Kỳ; 2. Sách – Lịch sử – Hoa Kỳ; 3. Thư viện – Lịch sử – Anh; 4. Sách – Lịch sử – Anh. Trong khi đó nếu dùng ngôn ngữ từ khóa ta chỉ thêm 2 từ khóa: Anh; Hoa Kỳ. Như vậy, vừa đơn giản, vừa gọn nhẹ và các từ khóa này không khác gì các đề mục chủ đề chính và phụ đề. Nhân đây, tôi cũng xin nhấn mạnh rằng chính cấu tạo như vậy nên bộ từ khóa, từ chuẩn sẽ có số lượng từ ít hơn rất nhiều so với ngôn ngữ từ khóa tiền kết hợp. Chẳng hạn, trong bảng phân loại UDC (Bảng phân loại thập phân quốc tế) mục 5 và 6 (Toán học. Các khoa học tự nhiên. Các khoa học ứng dụng. Y tế. Kỹ thuật) có gần 100.000 các phân mục (có nghĩa là đơn vị từ vựng của ngôn ngữ tìm tin) và gần 70.000 đề mục và tiêu đề mục trong bảng tra chữ cái - chủ đề thì trong ngôn ngữ từ khóa, từ chuẩn chỉ có 5 - 7 nghìn từ và tổ hợp từ. Hiện nay, Bộ đề mục chủ đề của Thư viện Quốc hội Mỹ dày tới 5.000 trang (có tài liệu nói là hơn 7.000 trang), đủ thấy nó sẽ gây khó khăn cho người sử dụng như thế nào.

Số từ khóa, cả tự do lẫn kiểm soát, định cho mẫu tìm một tài liệu là không xác định, từ một đến vài chục, do đó sự mô tả tài liệu sẽ đầy đủ hơn. Xử lý tài liệu bằng từ khóa sẽ giải phóng công bộ xử lý thụng tin khỏi các khuôn mẫu cứng nhắc của khung phân loại, các bảng đề mục chủ đề. Đồng thời cũng tránh được tính chủ quan của người xử lý tài liệu, các từ khóa được chọn có tính chất hình thức. Vì vậy việc định từ khóa, nhất là từ khóa tự do, có thể thực hiện được bằng cách tự động hóa. Tuy nhiên, ngôn ngữ từ khóa, kể cả từ khóa có kiểm soát không thể triệt tiêu hết hiện tượng đồng âm khác nghĩa, đồng nghĩa khác âm, dẫn tới hiện tượng nhiễu tin. Mặt khác, khi sử dụng ngôn ngữ từ khóa đòi hỏi phải có các phương tiện cơ giới hoá và tự động hoá chuyên dùng (chẳng hạn, ngày nay là máy tính).

4. So sánh khả năng tìm tin giữa ngôn ngữ từ khóa và ngôn ngữ chủ đề

Đối với ngôn ngữ đề mục chủ đề:

- Tìm tin theo chuyên đề lớn sẽ đầy đủ và chính xác hơn. Chẳng hạn như đề mục chủ đề *Giao thông* đã nêu ở trên.

- Tìm tin theo yêu cầu đa chuyên đề sẽ rất khó với đề mục chủ đề.

Đối với ngôn ngữ từ khoá, từ chuẩn.

- Cho phép thực hiện việc tìm tin có hiệu quả theo những kết hợp bất kỳ, không định trước của các dấu hiệu (do tách từ, ghép từ). Trong khi tìm tin, nếu một lệnh tìm nào đó mà không tìm được tư liệu thì có thể bớt một hay vài từ khóa. Lúc đó thì kết quả tìm sẽ rộng hơn nhưng rất có thể có những tài liệu cần cho người tìm. Trong trường hợp, tài liệu về một yêu cầu nào đó sẽ nhiều thì ta có thể thêm những từ khóa khác, hạn chế lệnh tìm.

- Việc tìm tin theo chuyên đề bằng từ khóa có những độ “ồn” tin nhất định. Chẳng hạn, khi có yêu cầu tìm những tài liệu về xuất khẩu ô tô từ Nhật bản sang Hoa Kỳ ta sẽ tìm theo các từ khóa: ô tô, xuất khẩu, Nhật bản, Hoa Kỳ. Như vậy, ở đây ta sẽ có cả những tài liệu về xuất khẩu ô tô của Mỹ, đồng thời, không rõ ai là người xuất khẩu ô tô hay nói cách khác cả Mỹ và Nhật bản đều là các nước xuất khẩu ô tô. Vì vậy, ngoài các từ điển thì khi tìm tin theo từ khóa hoặc từ chuẩn người tìm phải tạo ra một ngữ pháp riêng. Người sử dụng chỉ cần nắm được những kỹ năng trong sử dụng toán tử Boole.

Cả đề mục chủ đề lẫn từ chuẩn đôi khi tạo nên sự mô tả không thật phù hợp với chủ đề nội dung của tài liệu, gây mất tin. Người dùng tin nhiều khi lại tìm tin theo những thuật ngữ do tác giả tài liệu sử dụng chứ không theo những tổ hợp từ đã định trước bởi những người xử lý, mặc dù gần với tài liệu họ cần nhưng có khi lại mô tả chưa thật đúng về tài liệu họ cần. Trong trường hợp đó họ lại phải dựa vào từ khóa tự do.

Nhà khoa học Mỹ, F.Jonker cho rằng, nếu thông tin có mức độ tản mạn không lớn, không đòi hỏi áp dụng quá một - hai tiêu chuẩn khác nhau (các khía cạnh) của việc đánh chỉ số, và nếu những tiêu chuẩn đó luôn luôn chỉ có một cấp, thì về mặt lý thuyết có thể sử dụng hoặc khung phân loại thư mục - thư viện, hoặc các ngôn ngữ chủ đề. Khi thông tin có tính tản mạn lớn, thì khả năng duy nhất là sử

dụng loại ngôn ngữ tìm tin bằng từ khóa, từ chuẩn, bởi vì chỉ khi đó mới có thể đánh chỉ số theo bất kỳ một tập hợp đặc trưng nào và tới bất kỳ độ sâu nào.

Như vậy, việc tạo ra phương pháp xử lý và tìm tin hậu kết hợp là một bước tiến quan trọng. Ngôn ngữ từ khóa hiện được sử dụng rộng rãi trong việc tạo lập các CSDL thư mục của các cơ quan thông tin - thư viện trong nước và trên thế giới. Ngày nay, đánh một từ khóa có nghĩa bất kỳ trên internet ta cũng nhanh chóng nhận được những thông tin, những nguồn tài liệu có liên quan. Có thể khẳng định rằng ngôn ngữ từ khóa hoàn toàn thay thế được ngôn ngữ chủ đề trong xử lý và tìm tin tư liệu ở bất cứ dạng tài liệu nào.

TÀI LIỆU THAM KHẢO

1. Lê Ngọc Oánh. Đào tạo các cán bộ thư viện phụ trách công tác thiết lập tiêu đề đề mục trong các thư viện//Tạp chí Thư viện Việt Nam. – 2008. – Số 4. – Tr. 3- 15.
2. Lê Ngọc Oánh. Những nguyên tắc định tiêu đề đề mục (Subject heading)//www.hcmuns.edu.vn.
3. Mikhailov, A.I. Cơ sở thông tin học/A.I. Mikhailov, A.I. Treynyi, R.C. Giliarepxkii. – M.: khoa học, 1968. – 756 tr. (tiếng Nga).
4. Nguyễn Minh Hiệp. Bản chất của công việc định chủ đề//Tạp chí Thư viện Việt Nam. – 2006. – Số 3. – Tr. 26- 32.
5. Nguyễn Minh Hiệp. Định chủ đề với ngành biên mục học hiện nay/Thư viện – công nghệ thông tin. – 2008. Số 5. – Tr. 2 – 9.
6. Thông tin học hiện đại: khoa học, công nghệ, hoạt động/R. X. Giliarepxki, G.Z. Zalaev... – M., 1997. – 212 tr. (tiếng Nga)