Quy Trình Thiết Kế hệ Thống Tìm Tin Trong Thông Tin - Thư Viện

Tổ chức hệ thống tra cứu thông tin thư viện

Hệ thống tìm tin được biên soạn nhằm cung cấp những kiến thức cơ bản về hệ thống tìm tin cho sinh viên chuyên ngành thông tin-thư viện và trở thành tài liệu tham khảo bổ ích cho các cán bộ thư viện.

1. Tổng quan về hệ thống tìm tin

1.1 Tìm tin

- ***Khái niệm tìm tin***: Tìm tin như một quá trình truyền thông một cách gián tiếp giữa các tác giả hoặc những người tạo lập các biểu ghi với những người sử dụng thông tin. Các ngôn ngữ và các kênh của hệ thống truyền thông này khác với các hệ thống truyền thông khác như truyền thông đại chúng hoặc truyền thông trực tiếp. Các ngôn ngữ được sử dụng trong hệ thống truyền thông này có thể là các ngôn ngữ tìm tin và/hoặc ngôn ngữ tự nhiên. Còn các kênh truyền thông có thể là các công cụ tìm tin như hệ thống mục lục, bảng tra, cơ sở dữ liệu…Nói cách khác, tìm tin là một quá trình tương tác giữa người sử dụng và các mảng tin thông qua các công cụ tìm tin khác nhau.
- ***Quá trình tìm tin***: Tìm tin là một quá trình cơ bản của con người và nó liên quan mật thiết với việc học tập và giải quyết vấn đề. Quá trình tìm tin được bắt đầu với nhu cầu tin của người sử dụng. Để đạt được mục tiêu như giải quyết một vấn đề, trả lời một câu hỏi cụ thể hoặc để thỏa mãn tính ham hiểu biết, người dùng tin có thể cần thông tin nhanh và ngắn gọn hoặc thông tin đầy đủ và chi tiết.
- **Quá trình tìm tin** là một quá trình tương tác phụ thuộc vào khả năng của người dùng tin, sự phản hồi từ hệ thống tìm tin và các quyết định của người dùng tin về các hành động tiếp theo dựa trên sự phản hồi này. Các chi tiết về nhu cầu tin ban đầu của người sử dụng có thể thay đổi. Các nhu cầu tin ban đầu thường được điều chỉnh sau khi người tìm tin biết nhiều hơn về vấn đề đang tìm kiếm thông qua sự tương tác với các hệ thống tìm tin.

Vì vậy, quá trình tìm tin tiếp tục đến khi người dùng tin có được thông tin thỏa mãn nhu cầu tin đã được điều chỉnh của mình. Công nghệ thích hợp, chẳng hạn hệ thống tìm tin và giao diện người sử dụng thích hợp, có thể thúc đẩy quá trình nhưng đó không phải là vấn đề cơ bản nhất vì quá trình tìm tin phụ thuộc nhiều vào người dùng tin và nhu cầu tin của người dùng tin cũng như bản chất, số lượng và sự đa dạng của thông tin.

1.2 Các dạng tìm tin

Có thể phân chia các dạng tìm tin theo các tiêu chí khác nhau như dựa vào tính chất của thông tin được tra cứu, dựa vào công cụ tìm tin được sử dụng, dựa vào loại hình tài liệu, dựa vào thời gian xuất bản của tài liệu, dựa vào ngôn ngữ tài liệu... Trên thực tế, để tìm tin một cách hiệu quả, quá trình tìm tin thường được thực hiện dựa trên sự kết hợp nhiều dạng tìm tin với nhau. Dưới đây là hai cách phân chia các dạng tìm tin thường được sử dụng.

Dựa vào tính chất của thông tin được tra cứu

Có thể phân chia thành các dạng tìm tài liệu và tìm thông tin dữ kiện.

Tìm tài liệu là quá trình xác định và chọn lọc các tài liệu từ các nguồn tìm tương ứng với yêu cầu tin hoặc các dấu hiệu tìm tin cho trước như tên tác giả, tên tài liệu, nơi xuất bản, nhà xuất bản…

Tìm thông tin dữ kiện là quá trình xác định, chọn lọc và tách ra khỏi nguồn tin những số liệu, dữ kiện cụ thể như các số liệu thống kê, các đặc tính, thông số kỹ thuật của các thiết bị, vật liệu, các khái niệm khoa học…để đáp ứng các yêu cầu tin.

Dựa vào công cụ tìm tin

Có thể chia thành các dạng tìm tin thủ công, bán tự động và tự động hóa.

Tìm tin thủ công là quá trình tìm tin dựa trên các công cụ tìm tin thủ công hay còn gọi là công cụ tìm tin truyền thống như hệ thống mục lục, bộ phiếu tra cứu, bảng tra, thư mục, ấn phẩm thông tin, tài liệu tra cứu…

Tìm tin bán tự động là quá trình tìm tin dựa trên các công cụ tìm tin bán tự động như phiếu lỗ mép và phiếu lỗ soi. Tìm tin tự động hóa là quá trình tìm kiếm các thông tin được xử lý, lưu trữ và truy cập thông qua máy tính hoặc mạng máy tính. Trước đây, tìm tin thủ công là dạng tìm tin phổ biến nhất trong các thư viện và cơ quan thông tin.

Tìm tin tự động hóa ngày càng phổ biến hơn và trở thành dạng tìm tin chủ yếu trong các thư viện và cơ quan thông tin lớn. Bên cạnh đó, các công cụ tìm tin bán tự động hầu như không còn được sử dụng nữa nên dạng tìm tin bán tự động ngày càng được ít người biết đến.

1.3 Hệ thống tìm tin

Hệ thống tìm tin (HTTT) được đề cập đến trong nhiều tài liệu khác nhau. Sau đây là một số định nghĩa về hệ thống tìm tin.

Một số khái niệm liên quan

Nhu cầu tin: Là nhu cầu khách quan của người dùng tin về những thông tin cần thiết cho công việc cụ thể của mình.

Yêu cầu tin: Là nhu cầu tin của người dùng tin được thể hiện dưới dạng văn bản hoặc lời.

***Mảng tin* (Information retrieval file):** Là tập hợp các tài liệu, dữ kiện (hoặc các thông tin về chúng) được sắp xếp theo một trình tự nhất định tiện lợi cho việc tìm và xử lý tin.

Mẫu tìm của tài liệu: Là nội dung cơ bản của tài liệu được thể hiện bằng các thuật ngữ của ngôn ngữ tìm tin. Mẫu tìm của một tài liệu được tạo lập trong quá trình xử lý tài liệu và được sử dụng để tìm tài liệu đó trong tập hợp nhiều tài liệu khác. Tập hợp các mẫu tìm của tài liệu là một bộ phận không thể thiếu của mảng tin. Quá trình thể hiện nội dung cơ bản của tài liệu bằng mẫu tìm được gọi là quá trình đánh chỉ số.

Lệnh tìm: Là nội dung của yêu cầu tin được thể hiện bằng các thuật ngữ của ngôn ngữ tìm tin.

Điểm truy cập (access point) là một từ, cụm từ, mã số, tên gọi…được sử dụng để tìm thông tin trong một hệ thống tìm tin. Điểm truy cập có thể là giá trị của các thuộc tính (hình thức và/hoặc nội dung) của đối tượng được phản ánh trong hệ thống tìm tin.

Ví dụ, trong một hệ thống tìm tin tư liệu, điểm truy cập có thể là tên tác giả, nhan đề, đề mục chủ đề, từ khóa, ký hiệu phân loại…cho phép tìm kiếm và nhận dạng một biểu ghi thư mục. Các điểm truy cập này được người xử lý tài liệu lựa chọn khi tạo lập một biểu ghi.

Đánh chỉ số (Indexing): Là quá trình thể hiện nội dung tài liệu và/hoặc yêu cầu tin bằng ngôn ngữ tìm tin.

Tính thích hợp (Relevance): Là mức độ trùng hợp giữa nội dung tài liệu với yêu cầu tin. [8] Tính phù hợp Tính phù hợp (Pertinence) là mức độ trùng hợp giữa nội dung tài liệu và nhu cầu tin. [8] Tiêu chuẩn phù hợp ý nghĩa Tiêu chuẩn phù hợp ý nghĩa là tập hợp các qui tắc

nhằm xác lập một cách hình thức mức độ thích hợp của tài liệu với yêu cầu tin. Có thể chia các tiêu chuẩn phù hợp ý nghĩa thành hai nhóm chính là định lượng và logic. Loại thứ nhất sử dụng các tiêu chí định lượng để đánh giá mức độ giống nhau về nội dung giữa tài liệu và yêu cầu tin (hệ số tương thích). điều này cho phép sắp xếp kết quả tìm được theo trật tự giảm dần về mức độ thích hợp của tài liệu với yêu cầu tin.

1.4 Mục đích của hệ thống tìm tin

Mục đích chung của một hệ thống tìm tin là giảm tối đa chi phí của người sử dụng để tìm thông tin cần thiết. Chi phí tìm tin có thể được tính bằng thời gian một người sử dụng phải bỏ ra trong tất cả các bước của quá trình tìm tin cho đến khi có được tài liệu hoặc thông tin cần thiết. Người sử dụng có thể gặp nhiều trở ngại trong quá trình tìm tin. Vì vậy, mục đích của một hệ thống tìm tin là hỗ trợ tối đa để người sử dụng có thể tìm được thông tin cần thiết một cách nhanh chóng, đầy đủ và chính xác.

1.5 Chức năng của hệ thống tìm tin và yêu cầu đối với hệ thống tìm tin

Chức năng chính của một hệ thống tìm tin bao gồm:

- Phân tích nội dung các tài liệu: phân tích và trình bày nội dung chính của tài liệu bằng các ngôn ngữ thích hợp
- Tổ chức và lưu trữ thông tin một cách thích hợp để có thể tìm kiếm thông tin theo các yêu cầu tin của người sử dụng;
- Phân tích các yêu cầu tin của người sử dụng và thể hiện các yêu cầu tin ở dạng thích hợp với việc tìm kiếm trong hệ thống
- Tìm trong hệ thống và lựa chọn thông tin thích hợp với yêu cầu tin

1.6 Thành phần của hệ thống tìm tin

Thành phần của một hệ thống tìm tin cụ thể bao gồm:

- Các mảng tin bao gồm tài liệu, thông tin về tài liệu/siêu dữ liệu, dữ kiện;

- Các công cụ logic-ngữ nghĩa, bao gồm ngôn ngữ tìm tin, các qui tắc sử dụng ngôn ngữ tìm tin và các tiêu chuẩn phù hợp ý nghĩa;

- Các phương tiện kỹ thuật đảm bảo thực hiện các chức năng của hệ thống;

- Các yếu tố đảm bảo cho việc khai thác hệ thống như nhân sự, tài liệu hướng dẫn sử dụng

Các thành phần cơ bản của hệ thống tìm tin được gọi là các phân hệ. Việc phân chia thành các phân hệ rất cần thiết và hữu ích cho việc thiết kế cũng như mô tả cơ chế vận hành của hệ thống tìm tin. Có nhiều cách chia hệ thống tìm tin thành các phân hệ, trong đó hai cách thường được sử dụng nhất là phân chia theo loại yếu tố và phương tiện đảm bảo hoạt động của hệ thống và phân chia theo nguyên tắc chức năng.

2. Hệ thống công cụ xử lý ngữ nghĩa

2.1 Khái niệm

Hệ thống công cụ xử lý ngữ nghĩa: Là tập hợp các công cụ ngôn ngữ -logic và phương pháp được sử dụng để xử lý, trình bày, tổ chức và tìm kiếm thông tin trong hệ thống tìm tin.

Ngôn ngữ tìm tin: Là ngôn ngữ nhân tạo được dùng để mô tả nội dung tài liệu hoặc yêu cầu tin và để tìm tin

Ngôn ngữ tìm tin được xây dựng để khắc phục các hạn chế của ngôn ngữ tự nhiên trong việc diễn đạt thông tin và tìm kiếm thông tin, bao gồm:

- Có nhiều ngôn ngữ tự nhiên và mỗi ngôn ngữ đều có vốn từ vựng rất lớn, trong đó có nhiều từ không thể sử dụng để xử lý tài liệu và tìm tin;

- Ngôn ngữ tự nhiên có nhiều loại từ và các loại từ có giá trị thông tin khác nhau;

- Có nhiều từ đồng nghĩa, từ đồng âm và ý nghĩa của các từ có thể thay đổi theo ngữ cảnh;

Những hạn chế nêu trên có thể dẫn đến tình trạng vừa thừa vừa thiếu khi sử dụng ngôn ngữ tự nhiên để xử lý và tìm thông tin.

Để khắc phục các hạn chế trên, ngôn ngữ tìm tin phải đáp ứng các yêu cầu sau:

- Quan hệ ngữ nghĩa một-một: mỗi khái niệm phải được biểu đạt bằng một thuật ngữ và ngược lại, một thuật ngữ phải biểu đạt một và chỉ một khái niệm.

- Cú pháp được xây dựng chặt chẽ và nhất quán: chỉ có một cách biểu đạt các khái niệm - Có lực ngữ nghĩa mạnh: Lực ngữ nghĩa của ngôn ngữ tìm tin là khả năng phản ánh chính xác và đầy đủ nội dung của tài liệu và yêu cầu tin.

- Bảo đảm tính khách quan của người sử dụng: Chỉ diễn đạt đặc trưng khách quan của các sự vật, hiện tượng và các mối tương quan giữa chúng.

- Tính mở: bảo đảm khả năng chỉnh sửa và bổ sung ngôn ngữ. Ngôn ngữ tìm tin được xây dựng dựa trên hai thành phần cơ bản là từ vựng và cú pháp. Từ vựng của ngôn ngữ tìm tin là tập hợp các đơn vị từ vựng (hay còn gọi là yếu tố từ vựng) được sử dụng để mô tả nội dung tài liệu và/hoặc yêu cầu tin. đơn vị từ vựng là các từ hoặc ký hiệu được sử dụng để diễn đạt các khái niệm. Mỗi đơn vị từ vựng diễn đạt một khái niệm. Từ vựng là thành phần chính của các ngôn ngữ tìm tin và đóng vai trò rất quan trọng đối với các chuyên gia thông tin - thư viện và người dùng tin trong việc xử lý tài liệu và tìm thông tin.

Cú pháp của ngôn ngữ tìm tin là tập hợp các mối quan hệ giữa các đơn vị từ vựng, các qui tắc biểu thị các mối quan hệ đó và các qui tắc sử dụng các đơn vị từ vựng để mô tả thông tin.

2.2 Các loại ngôn ngữ tìm tin

Có hai loại NNTT tiền kết hợp là ngôn ngữ phân loại và ngôn ngữ đề mục chủ đề. đặc trưng chính của các NNTT tiền kết hợp là từ vựng có cấu trúc phân cấp một cách hệ thống và thường ở dạng một danh mục được định sẵn với các đơn vị từ vựng là các từ, cụm từ hoặc mã số. Khi đánh chỉ số tài liệu, người xử lý tài liệu có thể sử dụng các đơn vị từ vựng có sẵn hoặc kết hợp các đơn vị từ vựng với nhau theo những qui tắc nhất định để diễn tả các khái niệm phức tạp.

- **Ngôn ngữ phân loại**

Ngôn ngữ phân loại là ngôn ngữ tìm tin chuyên dụng cho phép người sử dụng tiếp cận tài liệu theo lĩnh vực tri thức được thể hiện trong nội dung tài liệu. Ngôn ngữ phân loại được sử dụng để phân loại tài liệu. Phân loại tài liệu là sự phân chia các tài liệu thành nhóm theo các dấu hiệu nhất định như lĩnh vực tri thức, vấn đề, đối tượng hoặc theo các dấu hiệu hình thức.

- **Ngôn ngữ tìm tin từ khóa**

Từ khóa là từ hoặc cụm từ ổn định, đơn nghĩa được sử dụng để mô tả nội dung chính của tài liệu và để tìm tin. Ngôn ngữ từ khóa là ngôn ngữ tìm tin hậu kết hợp có từ vựng được cấu thành từ các đơn vị từ vựng là từ khóa dựa trên ngôn ngữ tự nhiên, được sử dụng để xử lý tài liệu và yêu cầu tin.

- **Ngôn ngữ tìm tin đề mục chủ đề (ĐMCĐ)**

Ngôn ngữ ĐMCĐ là ngôn ngữ tìm tin có từ vựng là một tập hợp các từ hoặc cụm từ từ ngôn ngữ tự nhiên, được sử dụng để mô tả nội dung tài liệu và để tìm tin. Đề mục chủ đề là từ hoặc cụm từ được sử dụng để trình bày chủ đề của tài liệu hoặc yêu cầu tin. Từ vựng của ngôn ngữ tìm tin ĐMCĐ là bảng đề mục chủ đề. Bảng đề mục chủ đề là tập hợp các ĐMCĐ được sắp xếp theo vần chữ cái, đảm bảo sao cho các khái niệm được trình bày rõ ràng và không trùng lặp.

Ngôn ngữ mô tả tài liệu điện tử

Ngôn ngữ mô tả tài liệu điện tử

Ngôn ngữ HTML

Ngôn ngữ XML

Siêu dữ liệu (meta)

2.3 Tổ chức thông tin trong hệ thống tìm tin

Hệ thống tìm tin tư liệu là hệ thống tìm tin được phổ biến rộng rãi nhất trong các CQTT-TV. Vì vậy, chương này chủ yếu đề cập đến cách tổ chức thông tin trong các hệ thống tìm tin tư liệu.

Thành phần chính của một hệ thống tìm tin bất kỳ là các tập tin chứa thông tin về các thực thể được phản ánh trong hệ thống. Thực thể có thể là các đối tượng (như con người, tổ chức, tài liệu, vật liệu…) hoặc quá trình, hiện tượng trong thế giới khách quan mà con người có thể nhận dạng và mô tả được. Mỗi một thực thể được mô tả bằng một tập hợp các thuộc tính khác nhau (bộ thuộc tính).

Thuộc tính là những đặc trưng, tính chất phản ánh nội dung hoặc hình thức của thực thể mà con người có thể nhận dạng và trình bày được. Mỗi thuộc tính có một tên và một/nhiều giá trị hoặc nội dung. Nội dung có thể tương đương với giá trị của một thuộc tính hoặc chỉ đề cập một phần của giá trị. Giá trị là các ký tự hoặc bộ ký tự có ý nghĩa, được sử dụng để thể hiện nội dung các thuộc tính. Ví dụ, “ðHQG Tp.HCM” là một giá trị của thuộc tính “Nhà xuất bản” của tài liệu.

Trong trường hợp một thuộc tính có nhiều giá trị thì giá trị đó được gọi là giá trị lặp. Trường hợp một thuộc tính có thể nhận một trong hai giá trị (có hai giá trị khả dĩ) thì gọi là giá trị nhị phân. Thực thể được phản ánh trong hệ thống tìm tin tư liệu là tài liệu. Các tập tin trong hệ thống tìm tin tư liệu chứa thông tin về tài liệu – là tập hợp các giá trị của các thuộc tính hình thức và nội dung của tài liệu. Thuộc tính hình thức của tài liệu được thể hiện bằng các yếu tố như tên tác giả, nhan đề, các yếu tố xuất bản, dạng tài liệu, số ký hiệu…

Thuộc tính nội dung của tài liệu được thể hiện bằng các thuật ngữ của ngôn ngữ tìm tin được sử dụng trong hệ thống tìm tin. Tập hợp các giá trị của các thuộc tính nội dung chính là mẫu tìm. Quá trình tìm tin trong hệ thống tìm tin tư liệu là quá trình so sánh lệnh tìm với mẫu tìm của tài liệu.

Vì vậy, các mẫu tìm của tài liệu trong các tập tin phải được tổ chức sao cho việc so sánh giữa mẫu tìm và lệnh tìm có thể thực hiện một cách dễ dàng. Một tài liệu được xem là “tìm được” khi mẫu tìm của nó tương thích với lệnh tìm được nhập vào hệ thống. Sau khi tài liệu được xử lý, thông tin về tài liệu được tổ chức trong các tập tin và được lưu trữ trong bộ nhớ của hệ thống tìm tin.

Bộ nhớ là một hệ thống vật mang tin được sử dụng để ghi lại và lưu trữ thông tin theo thời gian nhằm mục đích tìm kiếm và cung cấp thông tin theo yêu cầu. Hệ thống này có thể là hệ

thống mục lục truyền thống, các bộ phiếu lỗ mép, phiếu lỗ soi, các thiết bị nhớ như băng từ, đĩa từ, đĩa quang …

Thành phần của bộ nhớ bao gồm các biểu ghi.

Biểu ghi của bộ nhớ trong hệ thống tìm tin là vật mang tin được sử dụng để lưu trữ các yếu tố thông tin về tài liệu.

Biểu ghi của bộ nhớ trong hệ thống tìm tin có thể là phiếu mô tả trong hệ thống mục lục thủ công, biểu ghi trong các tập dữ liệu… Có hai nguyên tắc lưu trữ thông tin trong hệ thống tìm tin tư liệu, bao gồm:

- Lưu trữ theo tài liệu: mỗi tài liệu tương ứng với một biểu ghi chứa mẫu tìm của tài liệu đó.

- Lưu trữ theo nội dung của tài liệu: mỗi thuật ngữ của NNTT thể hiện chủ đề của tài liệu (ví dụ một từ khóa/một ĐMCĐ) tương ứng với một biểu ghi trên đó liệt kê số ký hiệu của tất cả các tài liệu có nội dung đề cập đến chủ đề đó.

Nhìn chung, có thể có ba cách tổ chức các biểu ghi trong bộ nhớ của hệ thống tìm tin tư liệu, tương ứng với hai nguyên tắc lưu trữ thông tin nói trên như sau: Pi --> Di , Pi - -> ai , di - -> ai1, ai2, ai3, ..., ain Trong đó Pi - mẫu tìm của tài liệu; Di – tài liệu hoặc bản sao tài liệu; ai - số ký hiệu/địa chỉ lưu trữ tài liệu hoặc bản sao tài liệu; di - thuật ngữ của NNTT; ai1, ai2, ai3, ..., ain - các số ký hiệu/địa chỉ lưu trữ các tài liệu có chứa di trong mẫu tìm. Về bản chất, hai cách đầu tiên là hai trường hợp của cùng một sơ đồ tổ chức. Vì vậy, có thể có hai sơ đồ tổ chức thông tin trong bộ nhớ của hệ thống tìm tin tư liệu là sơ đồ tổ chức tuyến tính và sơ đồ đảo.

3. Thiết kế hệ thống tìm tin

3.1 Tổng quan về thiết kế hệ thống tìm tin

Thiết kế là khâu thiết yếu để phát triển hệ thống tìm tin và có tác động trực tiếp đến hiệu quả hoạt động của hệ thống. Quá trình thiết kế một hệ thống tìm tin bao gồm nhiều bước. Theo B.C.Vickery, khi thiết kế hệ thống tìm tin, người thiết kế phải xác định những vấn đề sau:

Các chức năng của hệ thống được thiết kế:

- Các mục tiêu của hệ thống;

- Các hệ thống lớn hơn có liên quan? Các chức năng và mục tiêu của các hệ thống này?Các khả năng thay đổi các mục tiêu và chức năng của hệ thống?

- Môi trường tổng thể của hệ thống;

- Các loại dịch vụ đầu ra của hệ thống và các đặc trưng của các dịch vụ này; - Các loại tài liệu đầu vào của hệ thống: đặc trưng và số lượng tài liệu được nhập vào hệ thống; - Các qui trình cần thiết để chuyển đổi đầu vào thành đầu ra theo dự tính; - Số lượng đầu vào, đầu ra và mức độ giao dịch dự tính;

- Những trở ngại dự kiến có thể ảnh hưởng đến việc thiết kế và vận hành hệ thống;

- Các lựa chọn để thực hiện các mục tiêu của hệ thống được thiết kế;

- Cách đánh giá hiệu quả hoạt động của hệ thống. Để thực hiện việc thiết kế một cách hiệu quả, người thiết kế một hệ thống tìm tin cần nắm được những thông tin sau:

- Các đặc điểm, qui mô và vị trí của các nhóm người sử dụng mục tiêu;
- Nhu cầu tin của các nhóm người sử dụng mục tiêu: nội dung, hình thức và mức độ thường xuyên của nhu cầu tin;
- Các yêu cầu đối với hệ thống và các dịch vụ được cung cấp;

- Các hệ thống hiện hữu có thể phát triển, thay thế hoặc cạnh tranh;
- Phạm vi bao quát của hệ thống: nội dung (các lĩnh vực được bao quát) và qui mô của vốn tài liệu; - Các qui trình lưu trữ và tìm tin thích hợp;
- Mức độ xử lý tài liệu: yêu cầu về độ sâu xử lý các tài liệu được nhập vào hệ thống; - Hình thức và loại hình đầu vào và đầu ra;
- Các yêu cầu đặc biệt như khả năng tương thích với các hệ thống khác, thiết bị cần thiết, kỹ năng của nhân viên…

3.2 Quy trình thiết kế hệ thống tìm tin

Quy trình thiết kế HTTT bao gồm các giai đoạn chính như sau:

- Xác định các mục tiêu và yêu cầu đối với hệ thống;

- Thiết kế cấu trúc tổng quát và xây dựng mô hình mẫu (Prototype) của hệ thống;

- Thử nghiệm mô hình mẫu;

- Hoàn chỉnh thiết kế hệ thống và vận hành hệ thống trên cơ sở các kết quả thử nghiệm;

- Kiểm tra, đánh giá hệ thống.

4. Hệ thống tìm tin thủ công

5.1 Hệ thống mục lục

Hệ thống mục lục thư viện (hay thường được gọi là mục lục) là một tập hợp có tổ chức các biểu ghi phản ánh vốn tài liệu của một kho tài liệu hay một bộ sưu tập nào đó.

Một bộ sưu tập có thể bao gồm một hoặc nhiều loại hình tài liệu như sách, ấn phẩm định kỳ, bản đồ, tranh ảnh, băng hình…Thông thường, mục lục phản ánh vốn tài liệu của một thư viện hoặc cơ quan thông tin.

Tuy nhiên, cũng có những mục lục phản ánh vốn tài liệu của nhiều thư viện, cơ quan thông tin nhằm hỗ trợ cho việc chia sẻ nguồn lực thông tin giữa các tổ chức như mục lục liên hợp.

Chức năng của mục lục: Mục lục có các chức năng cơ bản là nhận dạng, tập hợp, đánh giá hay chọn lọc và xác định vị trí của tài liệu. Các chức năng này có sự phụ thuộc lẫn nhau.

Chức năng nhận dạng hay tìm kiếm tài liệu: Mục lục được xây dựng nhằm tạo điều kiện cho người sử dụng có thể đối chiếu các dữ liệu về tài liệu đã biết với các biểu ghi trong mục lục để xác định thư viện có tài liệu đó hay không.

Chức năng tập hợp tài liệu: Chức năng này cho phép các biểu ghi của các tài liệu giống nhau hoặc có liên quan chặt chẽ với nhau về một phương diện hoặc dấu hiệu nào đó được tập hợp vào một chỗ trong mục lục

Chức năng đánh giá hay chọn lọc tài liệu: Chức năng này cho phép người sử dụng lựa chọn từ nhiều biểu ghi những tài liệu thích hợp nhất và tốt nhất chứa đựng kiến thức hoặc thông tin cần thiết.

Chức năng xác định vị trí tài liệu: Mục lục phản ánh địa chỉ lưu trữ tài liệu trong kho của một hoặc một số thư viện, cơ quan thông tin. Nhờ đó, người sử dụng có thể dễ dàng xác định được vị trí của tài liệu cần tìm.

Các hình thức mục lục thủ công

Mục lục sách

Mục lục sách: Là mục lục được ghi chép hoặc in dưới dạng các tập sách. Mục lục sách thường là các danh mục tài liệu được sắp xếp theo vần chữ cái tên tác giả, nhan đề hoặc

theo chủ đề. So với mục lục phiếu, mục lục sách có ưu điểm là có độ nén cao, cơ động, có thể đem theo để sử dụng ở mọi nơi và có thể xem lướt nhanh. Tuy nhiên, mục lục sách có hạn chế là chi phí cập nhật cao nên mục lục sách không được cập nhật thường xuyên như mục lục phiếu. Hiện nay mục lục sách vẫn được sử dụng cùng với các loại mục lục khác trong nhiều thư viện. Đặc biệt, mục lục lục sách vẫn được sử dụng như phương tiện duy nhất để truy cập các tài liệu quí hiếm trong một số thư viện và cơ quan lưu trữ.

Mục lục phiếu: Là mục lục được tạo thành từ những phiếu mô tả (biểu ghi) với kích thước khác nhau. Hiện nay, kích thước của phiếu chuẩn là 7,5 x 12,5 cm. Các phiếu mô tả có thể được chép tay, đánh máy hoặc được in từ máy tính. Bên cạnh những hạn chế như thiếu tính gọn nén, chiếm nhiều diện tích, không cơ động, mục lục phiếu cũng có các ưu điểm như dễ cập nhật và cho phép nhiều người sử dụng cùng một lúc.

Hiện nay mục lục phiếu vẫn được sử dụng trong nhiều thư viện. Có nhiều lý do để các thư viện duy trì mục lục phiếu như thư viện không có khả năng và điều kiện trang bị phần mềm và thiết bị cần thiết cho mục lục điện tử; thư viện không có khả năng chuyển đổi hồi cố hoàn toàn mục lục phiếu sang mục lục điện tử hoặc đơn giản là do nó phù hợp với thói quen của người sử dụng thư viện.

Các thành phần của hệ thống mục lục

Một hệ thống mục lục truyền thống gồm ba thành phần là mục lục công cộng, mục lục công vụ và hộp phiếu tiêu đề chuẩn.

Mục lục công cộng (public access catalog): Là mục lục dành cho người sử dụng truy cập tự do. Tùy theo cách sắp xếp các phiếu mô tả, có thể chia mục lục công cộng thành mục lục chữ cái, mục lục phân loại, mục lục chủ đề. Mục lục công vụ là mục lục sử dụng nội bộ (thường là mục lục vị trí, tổng mục lục chữ cái) phục vụ cho hoạt động nghiệp vụ của nhân viên thư viện. Mục lục vị trí phản ánh tổ chức kho tài liệu của một thư viện. Các phiếu mô tả trong mục lục được sắp xếp theo trật tự của tài liệu trên giá. Vì vậy, mục lục vị trí còn được gọi là mục lục xếp giá. Các phiếu mô tả của mục lục vị trí thường phản ánh đầy đủ số lượng bản/tập và vị trí của các tài liệu trong kho, đặc biệt khi một thư viện có nhiều kho tài liệu. Mục lục vị trí thường được bảo quản và sử dụng nội bộ ở bộ phận tổ chức kho.

- **Hệ thống mục lục chữ cái (MLCC)**

Mục lục chữ cái là hệ thống mục lục trong đó các phiếu mô tả được sắp xếp theo trật tự chữ cái tên tác giả hoặc nhan đề tài liệu được phản ánh trong mục lục.

Mục lục chữ cái là hệ thống tìm tin được tổ chức theo sơ đồ tuyến tính. Trong mục lục chữ cái, thông tin về mỗi tài liệu được trình bày trên một phiếu mô tả.

Các phiếu mô tả được sắp xếp dựa trên các đặc trưng hình thức của tài liệu là tên tác giả/nhan đề tài liệu.

Cách sắp xếp này cho phép người sử dụng có thể trả lời được hai câu hỏi cơ bản sau:

- Thư viện có hay không một tài liệu cụ thể với tên tác giả và/hoặc nhan đề đã được biết trước;

- Thư viện có những tác phẩm nào của một tác giả cụ thể.

Cấu trúc mục lục chữ cái

Các thành phần của mục lục chữ cái bao gồm các phiếu mô tả, các phiếu tiêu đề và các phiếu chỉ chỗ.

Phiếu mô tả trình bày các yếu tố đặc trưng cơ bản của một tài liệu dưới một hình thức chuẩn hóa giúp người sử dụng nhận dạng và phân biệt tài liệu này với các tài liệu khác. Phiếu mô tả có kích thước theo qui định là 7,5 x 12,5cm .

Phiếu chỉ chỗ 63 để kiểm soát tính thống nhất trong hệ thống mục lục nhằm tạo điều kiện thuận tiện cho việc sử dụng, bên cạnh các phiếu mô tả mục lục chữ cái còn có các phiếu chỉ chỗ (cũng được xếp theo vần chữ cái).

Cách tổ chức phiếu mô tả trong mục lục chữ cái

Mục lục chữ cái kiểu từ điển

Mục lục chữ cái kiểu phân đoạn

Qui tắc tổ chức chung nhất là tất cả các phiếu (bao gồm phiếu mô tả, phiếu tiêu đề, phiếu chỉ chỗ) trong MLCC đều được xếp theo trật tự vần chữ cái của từng ngôn ngữ hay của từng hệ ngôn ngữ được chọn để thể hiện tiêu đề mô tả.

Trong một số trường hợp, các phiếu được sắp xếp không theo nguyên tắc này. Ví dụ, các tác phẩm của cùng một tác giả được xếp theo thứ tự lần lượt là toàn tập, tuyển tập và các tác phẩm riêng lẻ (ở phần này các tác phẩm được xếp theo trật tự chữ cái tên tác phẩm).

5.2 Hệ thống mục lục phân loại

Mục lục phân loại (MLPL) là hệ thống mục lục trong đó các phiếu mô tả được sắp xếp theo các môn loại tri thức dựa trên một hệ thống phân loại nào đó. Mục lục phân loại phản ánh kho tài liệu của thư viện theo nội dung các ngành khoa học một cách hệ thống.

Ưu điểm của mục lục phân loại là không bị ảnh hưởng bởi sự thay đổi thuật ngữ vì nó sử dụng các ký hiệu, chữ và/hoặc số thay cho từ ngữ.

Nhược điểm lớn nhất của MLPL là do được xây dựng dựa trên một khung phân loại nhất định nên nếu người dùng tin không quen với các ký hiệu phân loại thì sẽ gặp khó khăn khi tra cứu và nhiều khi cần có sự hướng dẫn của nhân viên thư viện.

Các bộ phận cấu thành phần MLPL bao gồm các phiếu mô tả, các phiếu tiêu đề, các phiếu chỉ chỗ và phiếu ngăn.

Phiếu mô tả: Các yếu tố trong phiếu mô tả được trình bày thống nhất theo một qui tắc mô tả thư mục nhất định. Ký hiệu phân loại của tài liệu phải được thể hiện trên phiếu mô tả. đối với những tài liệu có nội dung phản ánh nhiều vấn đề, nghĩa là có từ hai ký hiệu phân loại trở lên thì phải có phiếu bổ sung cho các ký hiệu phân loại.

Phiếu tiêu đề: Trong MLPL cũng sử dụng các phiếu tiêu đề có gờ nhô lên như trong MLCC. Các phiếu tiêu đề được chia thành nhiều cấp tương đương với các lớp trong khung phân loại. Các phiếu tiêu đề có gờ nhô ở giữa thường được dành cho các lớp phân chia thứ nhất (tiêu đề cấp 1), thứ hai (tiêu đề cấp 2), đôi khi cho lớp thứ ba (tiêu đề cấp 3).

Phiếu chỉ chỗ qua lại (“cũng xem”) hướng dẫn người sử dụng tìm thêm ở các mục khác có liên quan để mở rộng nội dung và phạm vi tra cứu. Phiếu chỉ chỗ đi (“xem”) hướng dẫn người sử dụng tìm tài liệu ở đề mục chính xác trong trường hợp tài liệu có nội dung được phản ánh ở nhiều mục nhưng chỉ được xếp ở một mục nhất định. Phiếu ngăn Các phiếu ngăn có kích thước bằng phiếu mô tả nhưng khác màu, được dùng để phân cách các phiếu mô tả có cùng ký hiệu phân loại nhưng khác nhau ở một số đặc điểm khác như ngôn ngữ, loại hình tài liệu…

Cách tổ chức phiếu mô tả trong MLPL

Sơ đồ tổ chức thông tin được áp dụng trong MLPL là sơ đồ đảo, nghĩa là các phiếu mô tả được sắp xếp theo kí hiệu phân loại thể hiện nội dung tài liệu.

Trong MLPL, các phiếu mô tả có cùng ký hiệu phân loại được xếp chung sau mỗi phiếu tiêu đề trên đó ghi ký hiệu phân loại và tên đề mục tương ứng trong khung phân loại. ðể thuận tiện cho việc tra cứu, sau mỗi phiếu tiêu đề chỉ nên xếp khoảng 50 phiếu mô tả.

Các phiếu mô tả trong từng mục được xếp theo từng nhóm ngôn ngữ và có phiếu ngăn giữa các nhóm ngôn ngữ. Trong mỗi nhóm ngôn ngữ, các phiếu mô tả được xếp theo thứ tự vần chữ cái các tiêu đề mô tả (tên tác giả hoặc nhan đề tài liệu).

Cũng có thể sắp xếp các phiếu mô tả theo thứ tự thời gian ngược của năm xuất bản và các lần xuất bản tài liệu (tài liệu mới xuất bản được xếp lên trước các tài liệu cũ hơn) để ưu tiên giới thiệu các tài liệu mới.

Mục lục phân loại được chứa trong các hộp phiếu trong tủ mục lục. Trên mỗi hộp phiếu có nhãn ghi ký hiệu phân loại và các mục chứa trong hộp phiếu.

Hộp phiếu tra chủ đề - chữ cái

Mục lục chủ đề: *L*à hệ thống mục lục trong đó các phiếu mô tả được sắp xếp theo chủ đề của tài liệu. Mục lục chủ đề phản ánh vốn tài liệu của thư viện theo chủ đề của nội dung tài liệu. Mục lục chủ đề đặc biệt hữu dụng khi phản ánh nội dung tài liệu thuộc các ngành khoa học ứng dụng.

Vì vậy, mục lục này thường được sử dụng trong các thư viện, cơ quan thông tin chuyên ngành. Mục lục chủ đề giúp người dùng tin tìm tài liệu cần thiết theo từng chủ đề cụ thể. ðối với nhân viên thư viện, mục lục chủ đề là công cụ hỗ trợ biên soạn các thư mục chuyên đề, tra cứu theo yêu cầu, tuyên truyền, giới thiệu tài liệu theo các chuyên đề…

Cấu trúc mục lục chủ đề

Phiếu mô tả trình bày đầy đủ các yếu tố mô tả như trong mục lục chữ cái. Ở phía trên góc bên trái của phiếu ghi ký hiệu xếp giá của tài liệu. Các đề mục chủ đề của tài liệu và phụ đề (nếu có) được liệt kê phía dưới các yếu tố mô tả. Nếu tài liệu có trên 2 đề mục chủ đề thì phải lập phiếu bổ sung cho các chủ đề.

Phiếu tiêu đề: Trong mục lục chủ đề cũng sử dụng các phiếu tiêu đề (bao gồm phiếu tiêu đề chính và phiếu tiêu đề phụ) để thể hiện các đề mục chủ đề, phụ đề và phân chia các phiếu trong cùng chủ đề. Thư viện có thể tự lập các phiếu tiêu đề hoặc dựa trên một bảng đề mục chủ đề có sẵn.

Phiếu chỉ chỗ: Phiếu chỉ chỗ “xem” chỉ từ đề mục chủ đề không thông dụng sang đề mục chủ đề thông dụng. Phiếu chỉ chỗ “cũng xem” chỉ dẫn người sử dụng đến các đề mục chủ đề khác có liên quan với chủ đề mà người sử dụng quan tâm.

Cách tổ chức phiếu trong mục lục chủ đề

Các phiếu tiêu đề được sắp xếp theo vần chữ cái của các đề mục chủ đề, không phân biệt chủ đề đơn hay phức. Trong từng đề mục chủ đề, các phiếu mô tả được sắp xếp theo trật tự chữ cái các tiêu đề mô tả. Các phiếu được chứa trong các hộp phiếu trong tủ mục lục. Trên mỗi hộp phiếu có dán nhãn ghi đề mục chủ đề ở đầu và cuối hộp phiếu.

Các bộ phiếu thư mục

Bộ phiếu tra cứu chính

Các bộ phiếu chuyên đề

Các bộ phiếu theo loại hình tài liệu

Các bộ phiếu dữ kiện

6. Hệ thống tìm tin tự động hóa

6.1 Khái quát về hệ thống tìm tin tự động hóa

Hệ thống tìm tin tự động hóa (HTTTTĐH) là hệ thống tìm tin sử dụng máy tính điện tử để thực hiện một số chức năng. Một hệ thống tìm tin tự động hóa bao gồm các thành phần cơ bản sau:

- Hệ thống công cụ logic- ngữ nghĩa;

- Các phương tiện kỹ thuật đảm bảo thực hiện các chức năng của hệ thống; - Các cơ sở dữ liệu;

- Các yếu tố và phương tiện đảm bảo vận hành hệ thống. Hệ thống công cụ logic- ngữ nghĩa bao gồm các công cụ xử lý hình thức và nội dung tài liệu như các quy tắc mô tả thư mục, các khổ mẫu trao đổi dữ liệu, các ngôn ngữ tìm tin, các quy tắc sử dụng các công cụ logic-ngữ nghĩa, các tiêu chuẩn phù hợp ngữ nghĩa và một số công cụ ngữ nghĩa khác.

Hệ thống công cụ-ngữ nghĩa đóng vai trò quan trọng trong việc xử lý, trình bày, tổ chức và tìm kiếm thông tin.

Phân loại hệ thống tìm tin tự động hóa: Có thể phân loại hệ thống tìm tin tự động hoá dựa trên các cơ sở khác nhau. Sau đây là cách phân loại dựa trên mục đích cung cấp dịch vụ, phương thức truy cập và thế hệ của hệ thống tìm tin tự động hóa.

- Phân loại theo mục đích cung cấp dịch vụ Theo mục đích cung cấp dịch vụ, có thể phân các hệ thống tìm tin tự động hóa thành hai loại là hệ thống tìm tin nội bộ và hệ thống tìm tin thương mại. Hệ thống tìm tin nội bộ được xây dựng nhằm phục vụ người sử dụng thuộc một tổ chức nhất định.

Hệ thống tìm tin thương mại cung cấp dịch vụ truy cập các CSDL hoặc ngân hàng dữ liệu cho người sử dụng dựa trên cơ chế thị trường. Người sử dụng thanh toán phí dịch vụ dựa trên mức độ sử dụng dịch vụ.

- Phân loại theo phương thức truy cập 73 Theo phương thức truy cập CSDL, có thể chia các hệ thống tìm tin tự động hóa thành hai loại là hệ thống tìm tin cục bộ và hệ thống tìm tin trực tuyến.

Hệ thống tìm tin cục bộ cung cấp khả năng truy cập các CSDL được lưu trữ trên máy tính điện tử tại chỗ trong khi hệ thống tìm tin trực tuyến cung cấp khả năng truy cập các CSDL được lưu trữ trên các máy tính điện tử ở xa thông qua mạng máy tính và các phương tiện viễn thông.

Phân loại theo thế hệ của hệ thống tìm tin: Theo thế hệ của hệ thống tìm tin, có thể chia các hệ thống tìm tin tự động hóa thành ba loại như sau:

- *Hệ thống tìm tin thế hệ thứ nhất*: Là hệ thống cung cấp khả năng tìm thông tin thư mục với các lệnh tìm ở dạng dòng lệnh. Cách sử dụng hệ thống phức tạp, đòi hỏi người sử dụng phải có những kỹ năng tìm tin nhất định. Vì vậy, người sử dụng hệ thống này chủ yếu là các chuyên gia thông tin.

- *Hệ thống tìm tin thế hệ thứ hai*: Là hệ thống cung cấp khả năng tìm thông tin thư mục và toàn văn với các lệnh tìm ở dạng hệ thống thực đơn. Cách sử dụng hệ thống này đơn giản hơn so với hệ thống thế hệ thứ nhất. Người sử dụng hệ thống có thể là các chuyên gia hoặc người sử dụng thông thường.

- *Hệ thống tìm tin thế hệ thứ ba*: là hệ thống cung cấp khả năng tìm thông tin đa phương tiện, cho phép người sử dụng khai thác thông tin một cách dễ dàng qua các giao diện đồ họa.

Các chức năng của hệ thống tìm tin tự động hóa

Chức năng tìm tin: Cho phép tìm các biểu ghi trong CSDL đáp ứng yêu cầu tin cụ thể của người sử dụng. Để thực hiện việc tìm tin trong hệ thống, người sử dụng phải thể hiện yêu cầu tin bằng biểu thức tìm.

Logic Bool cho phép người sử dụng kết hợp nhiều khái niệm với nhau một cách logic để xác định thông tin cần thiết. Các phép toán của logic Bool cho phép thực hiện ba loại quan hệ cơ bản giữa các khái niệm là quan hệ tương giao, quan hệ kết hợp và 74 quan hệ loại trừ.

Toán tử lân cận được sử dụng để qui định khoảng cách yêu cầu giữa hai thuật ngữ tìm trong một biểu ghi nhằm tăng độ chính xác của việc tìm tin.

Tìm cụm từ chính xác: Một cụm từ bao gồm hai hoặc nhiều từ và được xem như một đơn vị ngữ nghĩa độc lập. Ví dụ, “United States of America” là một cụm từ bao gồm bốn từ, được xác định như một thuật ngữ tìm thể hiện một khái niệm ngữ nghĩa cụ thể (một quốc gia) và có thể được sử dụng với một toán tử bất kỳ trong số các toán tử đã được đề cập ở trên.

Toán tử chặt từ (truncation): Kỹ thuật tìm tin với toán tử chặt từ cho phép mở rộng một thuật ngữ tìm bằng cách chặt bớt một phần của thuật ngữ và tìm tất cả các biểu ghi có chứa phần còn lại của thuật ngữ đó.

Kỹ thuật tìm so sánh: Kỹ thuật tìm so sánh cho phép tìm theo các dấu hiệu định lượng. Toán tử so sánh (=, , <=, >=) được sử dụng trong nhiều hệ thống tìm tin để tìm theo giá trị số hoặc dãy số (hoặc ngày tháng).

Kỹ thuật tìm so sánh: Kỹ thuật tìm so sánh cho phép tìm theo các dấu hiệu định lượng. Toán tử so sánh (=, , <=, >=) được sử dụng trong nhiều hệ thống tìm tin để tìm theo giá trị số hoặc dãy số (hoặc ngày tháng).

Kỹ thuật tìm so sánh: Kỹ thuật tìm so sánh cho phép tìm theo các dấu hiệu định lượng. Toán tử so sánh (=, , <=, >=) được sử dụng trong nhiều hệ thống tìm tin để tìm theo giá trị số hoặc dãy số (hoặc ngày tháng).

Hiển thị kết quả

6.2. Cơ sở dữ liệu (CSDL)

Một số khái niệm cơ bản

Dữ liệu là một chuỗi các ký hiệu cơ bản như các số nguyên hoặc các ký tự và là giá trị của một thuộc tính.

Cấu trúc dữ liệu là thuật ngữ được dùng để chỉ một sơ đồ tổ chức các phần tử dữ liệu /thông tin có liên quan với nhau.

Hệ quản trị CSDL là tập hợp các chương trình hỗ trợ người sử dụng quản lý và khai thác CSDL với các chức năng cơ bản là mô tả dữ liệu, cập nhật dữ liệu và tìm kiếm dữ liệu

Phần mềm tư liệu là phần mềm thực hiện các chức năng quản lý, lưu trữ và tìm kiếm thông tin về tài liệu. Phần mềm tư liệu bao gồm nhiều mô đun thực hiện các chức năng khác nhau như thiết lập cấu trúc CSDL, cập nhật dữ liệu, tìm kiếm dữ liệu, hiển thị kết quả tìm trên màn hình và in kết quả tìm.

Có thể phân loại CSDL dựa trên nhiều cơ sở khác nhau. Sau đây sẽ đề cập các cách phân loại chính là theo tính chất dữ liệu, theo phạm vi bao quát đề tài và theo loại hình tài liệu.

Phân loại theo tính chất dữ liệu

Phân loại theo phạm vi bao quát đề tài

Xây dựng CSDL

Một hệ thống tìm tin có thể chứa nhiều loại CSDL khác nhau nhưng các CSDL phổ biến nhất trong các TV-CQTT là CSDL thư mục và CSDL dữ kiện. Hiện nay có những phương pháp kết hợp cho phép xây dựng cả hai loại CSDL này trên một cấu trúc CSDL tích hợp. Mục này chỉ đề cập cách xây dựng CSDL trong các hệ thống tìm tin với các phần mềm tìm tin văn bản. Xây dựng CSDL nằm trong giai đoạn 2 của qui trình thiết kế một hệ thống tìm tin tự động hóa. Nhiều quyết định quan trọng phải được đưa ra khi thiết kế CSDL và hoạt động của hệ thống tìm tin được xây dựng sẽ phụ thuộc nhiều vào những quyết định này.