THƯ VIỆN SỐ
KHÁI NIỆM VÀ THÁCH THỨC

Đỗ Quang Vinh

1. Mở đầu

Thư viện số là một trong năm hướng nghiên cứu chính về công nghệ thông tin ở Mỹ và trên thế giới hiện nay.

Thư viện số đã trở thành một lĩnh vực nghiên cứu tích cực bao gồm lưu trữ khối và các cơ chế truy cập từ xa, cũng như tổ chức và tìm kiếm thông tin lưu trữ điện tử. Những đề xuất mới đối với thư viện số tiếp cận tới lưu trữ sách, báo, tạp chí định kỳ, bằng sáng chế, hồ sơ y học, sách hướng dẫn và v.v.... Trong nhiều phạm trù, thành công của các đề xuất này thuộc về cách các tài liệu lưu trữ được phân loại và cách thông tin này được sử dụng khi tìm kiếm chúng. Trong ngữ cảnh này, thông tin mô tả về một nguồn tin được gọi là siêu dữ liệu. Trong số khác, hầu hết siêu dữ liệu thông thường trợ giúp bởi các hệ thống hiện thời là tác giả, nhan đề, nhà xuất bản, chủ đề, ngày tháng, kiểu, nguồn tin, người đóng góp, vai trò, ISBN và v.v... Các đầu mục này được lưu trữ một lần cùng với thông tin tham chiếu tới, chúng được chỉ số hóa và sử dụng trong khi tìm kiếm tài liệu.

Trong bài báo này, chúng tôi trình bày tổng quan về thư viện số, một số kỹ thuật nổi bật đối với xây dựng thư viện số cỡ lớn, xu hướng phát triển của thư viện số và những thách thức cần phải giải quyết khi nghiên cứu thư viện số trong tương lai.

2. Khái niệm

Thư viện số là một thực thể liên quan tới sự tạo ra các nguồn tin và sự hoạt động thông tin qua các mạng toàn cầu. Một thư viện số được biểu thị là một tập hợp các máy chủ tự phân tán làm việc đồng thời để trao cho khách hàng diện mạo của một tập hợp liên kết đơn. Trong thực tế, mỗi máy chủ lưu trữ một lượng lớn thông tin đa dạng trên nhiều loại vật tải lưu trữ. Các cá nhân truy cập thông tin sẽ có một dải rộng chuyên môn trong những lĩnh vực liên quan tới truy cập khoá, như là học vấn máy tính, khả năng điều hướng kho tài liệu và tri thức lĩnh vực.

Đặc điểm của thư viện số là trợ giúp cộng tác, bảo quản tài liệu số, quản trị cơ sở dữ liệu phân tán, siêu văn bản, lọc thông tin, tìm kiếm thông tin, các đơn thể hướng dẫn, các quyền sở hữu trí tuệ, các dịch vụ thông tin multimedia, trả lời câu hỏi và các dịch vụ tra cứu, khám phá tài nguyên và phổ biến thông tin có chọn lọc. Chúng cho phép thông tin được truy cập toàn cầu, sao chép không lỗi, lưu trữ cô đặc và tìm kiếm nhanh.

2.1 Các thành phần chính

2.1.1 Hệ quản lý nội dung

Hệ quản lý nội dung là trung tâm của thư viện số. Không có nội dung số, sẽ không có thư viện số. Hệ quản lý nội dung bao hàm tập hợp tất cả chức năng thực hiện nhằm tạo ra một loại nội dung cụ thể, chẳng hạn tạp chí điện tử cho người dùng cuối. Một hệ quản lý nội dung có hai thành phần chính: hệ truy cập thông tin và hệ quản lý thông tin.

2.1.1.1 Hệ truy cập thông tin

Hệ truy cập thông tin có giao diện người dùng thích hợp hơn. Sự truy cập thông tin quy về loại chức năng có thể có được cho sử dụng hệ thống. Nó bao gồm các chức năng thường cung cấp cho loại dữ liệu riêng biệt, chẳng hạn, trong trường hợp của dữ liệu địa lý là chức năng vẽ bản đồ.

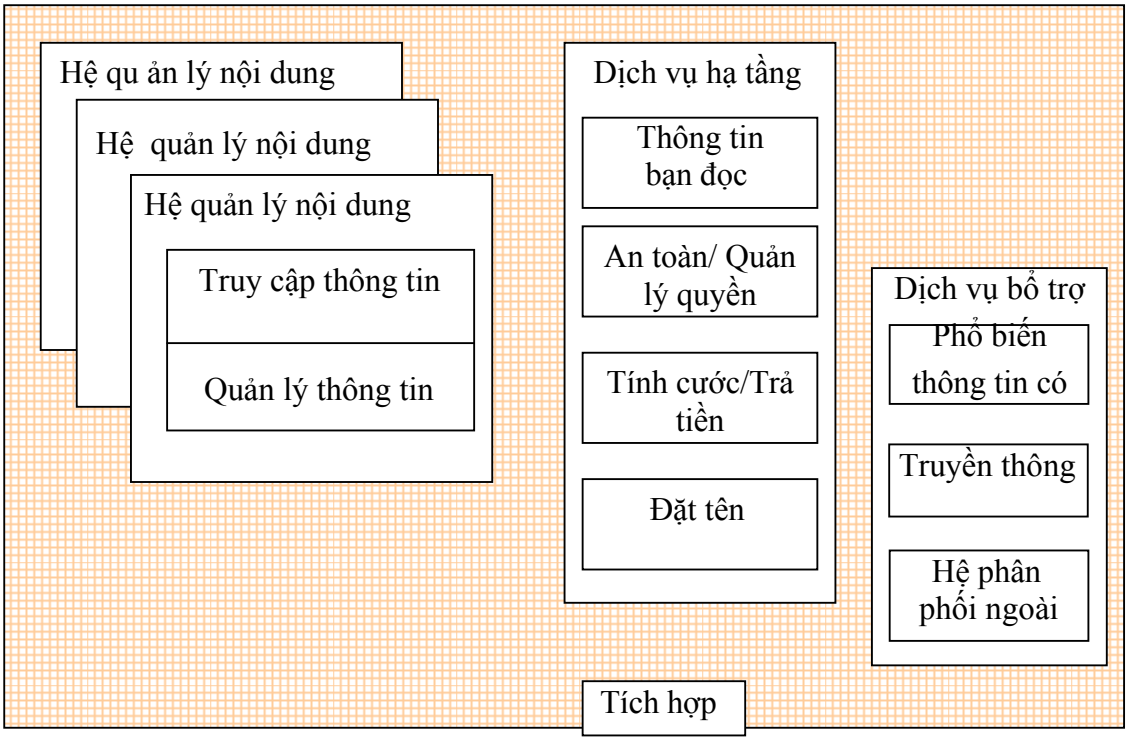
Truy cập thông tin bao hàm tìm kiếm, xem nội dung và xử lý thông tin. Một số loại thông tin cần phải xử lý sau khi tìm được. Chẳng hạn, tệp ảnh TIFF lớn (Target Image File Format) có thể cần được chuyển đổi thành tệp GIF (Graphics Interchange Format) được xem dễ dàng hơn với một trình duyệt Web.

2.1.1.2 Hệ quản lý thông tin

Quản lý thông tin cần phải làm cho truy cập thông tin là khả thi. Các chức năng truy cập thông tin cụ thể không thể có được nếu không có kiểu thích hợp về lưu trữ cơ bản và cơ chế quản lý, liệu có phải là một hệ cơ sở dữ liệu, động cơ tìm kiếm .v.v... Mỗi kiểu dữ liệu đòi hỏi hệ quản lý nội dung của riêng nó.

Quản lý nội dung không thể có được nếu không có thu thập nội dung. Thu thập được thực hiện bằng cách mua từ các nhà cung cấp, hoặc thông qua phát triển thư viện nội bộ như là quét tài liệu.

Hình 1: mô tả về các thành phần dịch vụ của thư viện số



❖ Các cách tiếp cận quản lý thông tin

Có bốn cách tiếp cận quản lý thông tin chính:

1. Các tệp độc lập với nội dung thư viện số có thể được bảo trì trên một máy chủ WWW hoặc FTP.

2. Một chương trình có thể được xây dựng để tự động hoá phần nào đó của tương tác với nội dung.
3. Quản lý nội dung với một động cơ tìm kiếm, như WAIS hoặc AltaVista.
4. Sử dụng một hệ quản trị cơ sở dữ liệu như SQL Server, Oracle, DB2 hoặc Informix.

Các cách tiếp cận là không loại trừ lẫn nhau; có nhiều cách kết hợp khác nhau để khai thác ưu thế riêng của từng kỹ thuật.

2.1.2 Dịch vụ hạ tầng

Bốn dịch vụ tạo thành hạ tầng quan trọng của một thư viện số: đặt tên, thông tin bạn đọc, an toàn và tính cước.

2.1.3 Dịch vụ bổ trợ

Ba dịch vụ bổ trợ chính ở thư viện số bao gồm: dịch vụ truyền thông, dịch vụ phân phối, dịch vụ phổ biến thông tin có chọn lọc

2.1.4 Tích hợp

Các thành phần của thư viện số phải được tích hợp. Một khi có nhiều hệ phần mềm cùng được sử dụng chúng ta cần phải quan tâm đến vấn đề liên kết chúng. Đây là một trong những thành phần kiến trúc phức tạp nhất của thư viện số.

Tích hợp bao hàm bài toán về cách làm cho hai hệ thống thông tin hoạt động đồng thời. Bài toán được giải quyết trong thư viện bằng cách chỉ cho bạn đọc sử dụng cả hai hệ thống, như một mục lục phân loại và một cơ sở dữ liệu trích dẫn.

Tích hợp là một bài toán thư viện số cơ bản. Nó xuất hiện ngay khi một thư viện quyết định cung cấp truy cập tới hai hệ thống thông tin khác nhau.

Các phương pháp tích hợp:

1. Các trang WWW kết nối nhiều Website.
2. Siêu dữ liệu Metadata.
3. Chuẩn tìm kiếm thông tin phân tán Z39.50.
4. CORBA (Comon Object Request Broker Architecture).

Không có một giải pháp nào là tốt hơn tất cả các giải pháp còn lại. Mỗi một giải pháp có vị trí riêng của mình. Tạo một trang Web là đơn giản. Sử dụng Z39.50 hạn chế người dùng tìm kiếm phân tán. CORBA là con đường tương lai. Nó là linh động hơn và cung cấp khả năng cho một môi trường tích hợp nhiều hơn Z39.50.

2.2 Tài nguyên

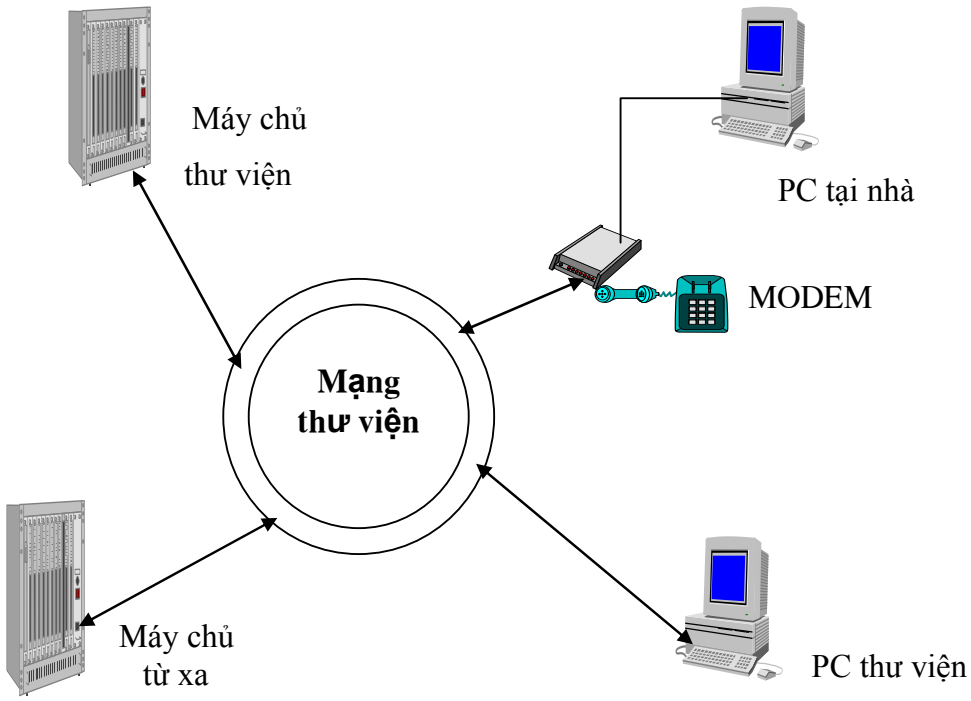
2.2.1 Phần cứng

Phần cứng là một thành phần chính, bao gồm: máy chủ, terminal và mạng. Có ba loại terminal: (1) máy tính gia đình, người sử dụng làm việc tại nhà và dùng modem để truy cập các dịch vụ; (2) terminal trong thư viện; (3) máy tính mạng. Tất cả terminal được kết nối tới máy chủ có thể đặt ở thư viện hoặc ở trường học qua mạng.

2.2.2 Đội ngũ

Đội ngũ của một thư viện số không chỉ là những kỹ thuật viên, nhà quản lý thư viện hệ thống và nhân viên thư viện dịch vụ điện tử, mà là tất cả cán bộ của một thư viện truyền thống.

Hình 2



3. Xây dựng thư viện số

Quá trình xây dựng một thư viện số điển hình bao hàm các pha sau:

- Tạo lập nội dung thư viện số
- Chỉ số hoá và lọc thông tin
- Trợ giúp truy cập phổ quát
- Bảo quản.

3.1 Tạo lập nội dung thư viện số

Thứ nhất, khi tạo lập một thư viện số, phải xác định loại thông tin cung cấp và loại thông tin loại bỏ. Không may, hầu hết thông tin thư viện số muốn cung cấp không được số hoá, nên yêu cầu bổ sung là xác định các quyền ưu tiên số hoá và các quá trình chuyển đổi.

Thứ hai là bản chất động của thông tin số hoá. Nội dung có thể thay đổi theo thời gian, đòi hỏi lưu trữ nhiều ấn bản, phải có cơ chế cho phép phân biệt ấn bản. Những thách thức liên quan khác bao gồm định danh các phương pháp bắt và chỉ số hoá vật tải liên tục ở thời gian thực và các kỹ thuật đối với xử lý, lưu trữ và quản trị lượng thông tin rất lớn và phức tạp.

Thứ ba, vì các liên kết siêu văn bản cho phép thư viện số cung cấp liên kết tới thông tin, thư viện số phải quyết định kiểu truy cập được cung cấp. Các vấn đề liên quan về quyền sở hữu và lưu trữ thông tin ngày càng trở nên quan trọng.

3.2 Chỉ số hoá và lọc thông tin

Sau khi vấn đề thu thập và lưu trữ được giải quyết, thách thức tiếp theo là tìm ra cách làm cho thông tin thích hợp có thể có đối với cá nhân có quyền tại thời điểm thích hợp. Các khách hàng phải có khả năng định danh, định vị và lọc thông tin sao cho chỉ thông tin thích hợp được đáp ứng và tổ chức nó thành các đơn vị quản lý được thông qua xếp hạng và phân loại. Tác nhân trí tuệ nhân tạo có thể được sử dụng nhiều để định vị và lọc thông tin trong tương lai.

Có hai loại quá trình định vị thông tin khác nhau:

- Loại thứ nhất là hữu ích trong tìm kiếm rộng, trong đó thông tin không cần được xác định cụ thể. Thông tin thích hợp bị tản mát nhiều trong một số nguồn tin không đồng nhất phân tán. Thách thức chính là biểu diễn tích hợp thông tin không liền mạch tới khách hàng. Sự lựa chọn là cung cấp các kỹ thuật tổ chức và phân loại có hiệu quả bó chùm thông tin thành đơn vị quản lý được.
- Loại thứ hai bao hàm tìm kiếm trọng tâm, hoàn toàn xác định và rất hẹp. Loại này đòi hỏi thông tin rất chi tiết. Vì tính chính xác là quan trọng nhất, các kỹ thuật lọc hiệu quả được dùng nhằm đáp ứng một lượng thông tin thích hợp nhất.

Mặt khác, giao diện người dùng là rất quan trọng. Thậm chí thông tin thích hợp nhất là không có giá trị, nếu khách hàng không hiểu sự trình bày. Những thư viện số tốt nhất là đồng dạng nhưng tuỳ biến, giao diện người dùng động có thể tích hợp trơn tru các kiểu dữ liệu thông thường từ nguồn có cấu trúc và không có cấu trúc với các kiểu dữ liệu đặc biệt (bản đồ, dữ liệu ba chiều và dữ liệu đồ hoạ liên tục). Những hệ thống này kết hợp các giải thuật và kỹ thuật cho phép tương hỗ ngữ nghĩa, sao cho người sử dụng có thể tìm kiếm ở lĩnh vực tri thức xa lạ bằng từ vựng và bản thể quen thuộc.

Vấn đề quan trọng khác của định vị thông tin là tìm ra quan hệ khoá, đặc biệt trong các nguồn tin không đồng nhất phân tán. Khai mỏ dữ liệu, trích lọc mẫu, liên kết và dị thường từ những nguồn dữ liệu lớn là lĩnh vực nghiên cứu đầy hứa hẹn, có thể tạo ra phần đáng kể đối với các ứng dụng thư viện số multimedia cỡ lớn phức tạp.

3.3 Trợ giúp truy cập phổ quát

Mục đích cuối cùng của một thư viện số là truy cập phổ quát, phù hợp với mục đích thư viện truyền thống là cung cấp truy cập thông tin công cộng. Nhằm thực hiện truy cập phổ quát, thư viện số cần phải giải quyết bài toán tích hợp thông tin và các nguồn tin không đồng nhất phân tán bằng cách thiết kế và cài đặt giao diện người dùng hiệu quả và giải quyết bài toán từ vựng.

Một trong những thách thức với cung cấp truy cập phổ quát là tạo ra các kỹ thuật trợ giúp tính đa dạng của thiết bị hiển thị thông tin trong điều khiển thông tin. Không chỉ có tính đa dạng của các hệ điều hành trong lĩnh vực máy tính, mà còn có tính đa dạng của các thiết bị hiển thị như tivi, máy fax, màn hình video và các thiết bị thông tin khác.

Thách thức chủ yếu khác là làm sao cho băng thông hạn chế có thể dùng được cho truyền thông tin thích nghi với số lượng người sử dụng tăng lên và kho dữ liệu lớn hơn, phức tạp hơn. Để đạt được truy cập phổ quát hợp lý, sự sử dụng thông minh băng thông, bao hàm khả năng bảo đảm băng thông cho một khoảng thời gian cho trước được định rõ và các chính sách trợ giúp sử dụng được ban hành.

3.4 Bảo quản

Vật tải điện tử không dễ phân huỷ như các loại khác. Tuy nhiên, sự thay đổi liên tiếp và sự nâng cấp khác ở dạng tài liệu điện tử (như chuẩn MARC, SGML, HTML, XML, .v.v...) và các vấn đề không tương thích cần phải được quan tâm chu đáo để bảo vệ dữ liệu.

4. Những áp lực thúc đẩy và xu hướng phát triển của thư viện số

Hơn 10 năm qua, có một số lợi ích khiến sự chuyển tiếp tới và sự mở rộng thư viện số được kiên trì ủng hộ. Bốn áp lực và xu hướng chính là: kinh tế, sự truy cập, công nghệ mới và các chuẩn.

4.1 Kinh tế

Sản xuất, lưu trữ, phân phối và tái sản xuất thông tin điện tử là rẻ hơn so với thông tin truyền thống. Hơn nữa, các thư viện số có thể hợp tác với nhau bằng cách cung cấp một cổng (liên kết) tới thông tin quản lý hoặc cung cấp bởi thư viện khác, cho phép chuyên môn hoá cũng như duy trì ngân sách thu thập và sản xuất trong khi vẫn cung cấp truy cập tới một lượng lớn thông tin. Những áp lực kinh tế khác hướng thư viện đến số hoá bao gồm:

- *Lạm phát*: sự tăng nhanh giá điều hành thư viện, đặc biệt ở thu thập hoặc mở rộng kho tài liệu các tạp chí học thuật. Trong 20 năm qua, giá tạp chí tăng vọt lên 400% trong khi giá sách và sách chuyên khảo tăng 40%.
- *Khối lượng*: sự bùng nổ về lượng, tính đa dạng và tính phức tạp của thông tin.
- *Bảo trì*: sự khủng hoảng bảo quản ở các kho tài liệu đang tồn tại, đặc biệt là giấy a xít.
- *Multimedia*: lượng thông tin multimedia tăng lên đòi hỏi các tiện ích xem hoặc nghe đặc biệt và các yêu cầu phân thành mục lục và lưu trữ khác.
- *Sự cộng tác*: những ưu điểm từ tài nguyên chia sẻ trong các thư viện và nhà cung cấp thông tin khác, cả về mặt kinh tế lẫn về cải thiện mức dịch vụ.
- *Truyền thông khoa học*: bài toán giá rất gay gắt vốn gắn với truyền thông khoa học truyền thống (như giá cung cấp truy cập quá mức tới số tạp chí học thuật thích hợp, tới bảo trì mức dịch vụ đầy đủ).

4. 2 Cải thiện mức dịch vụ

Thư viện số có khả năng cung cấp một mức dịch vụ không thể đạt được trước đây, nghĩa là sự tìm kiếm câu và từ riêng lẻ và phân phát thông tin tới người dùng, một thông tin không bị phân huỷ theo thời gian, dù nó là văn bản, âm thanh hoặc ảnh. Thông tin trước đây là không có sẵn hoặc khó thu thập, hiện nay thường là sẵn có về mặt điện tử. Truy cập thông tin có thể được cải thiện bằng một số cách: thời gian truy cập/tốc độ tìm kiếm, tính sẵn có, nội dung/tính thích đáng, tính trực quan đã cải thiện/giao diện người dùng. Về mặt lịch sử, sự nghiên cứu nâng cao chất lượng dịch vụ thư viện tập trung vào sự cải thiện nói chung tới truy cập thông tin. Xu hướng hiện tại là tuỳ biến riêng lẻ hoặc đáp ứng các phương pháp truy cập thông tin và giao diện người dùng.

Do đặc điểm sản xuất và phân phối thông tin điện tử, có những tiết kiệm đáng kể về thời gian từ sản xuất tới sử dụng. Thông tin điện tử chỉ cần tạo lập và lưu trữ một lần là ngay lập tức sẵn có trên mạng đồng thời tới nhiều người dùng, trái với nhiều bản sao chép được tạo ra theo thời gian và cung cấp qua các kênh phân phối truyền thống. Chẳng hạn, nhiều Website Internet mới đưa ra thông tin ở thời gian thực, không có thời gian trễ trong in ấn và phân phối.

4.3 Sử dụng công nghệ mới

Để đáp ứng có hiệu quả nhu cầu thông tin của khách hàng, thư viện số cần sử dụng một tổ hợp những thành tựu công nghệ và có khả năng thiết kế, xây dựng, quản trị và sử dụng các mạng điện tử toàn cầu. Nó phải có khả năng thích nghi nhanh với những thay đổi động trong công nghệ và đương đầu với kích cỡ, quy mô và tính phức tạp của các mạng lẫn thông tin có sẵn truyền qua chúng.

Nhiều thành tựu công nghệ trong sản xuất, quản trị và phân phối thông tin là nguyên nhân tạo khả năng cho thư viện số, bao gồm:

- Vật tải lưu trữ;
- Số hoá hoặc các kỹ thuật bắt thông tin (như công nghệ nhận dạng ký tự quang OCR);
- Chỉ số hoá tự động và tổ chức lượng thông tin lớn;
- Tốc độ tính toán;
- Công nghệ mạng (bao hàm nén dữ liệu);
- Tìm kiếm và phục hồi dựa trên nội dung;
- Tìm kiếm và phục hồi dựa trên đặc tính hoặc dựa trên kết cấu;
- Chỉ số hoá toàn văn;
- Khai phá tri thức hoặc tài nguyên;
- Multimedia và siêu văn bản hypertext;
- Các chuẩn: SGML (Standard General Markup Language), HTML (Hypertext Markup Language) và Z39.50;
- Các kỹ thuật hướng đối tượng;
- Cải tiến trong thiết kế giao diện người dùng và trực quan dữ liệu.

4.4 Các chuẩn

Để thư viện số thực sự là cổng toàn cầu, quan trọng là có các chuẩn kỹ thuật được thế giới chấp nhận đối với biểu diễn, tạo dạng, truyền thông tin và các giao thức. Đây là cách duy nhất đảm bảo tính tương thích. Vì thế, tính tương thích giữa thiết bị, dữ liệu, thực hành và thủ tục là cần thiết nhằm đạt được truy cập phổ quát và trao đổi thông tin điện tử toàn cầu. Không may, có nhiều rào cản xã hội, văn hoá và chính trị trước phát triển các chuẩn quốc tế, ngay cả khi lợi ích là rõ ràng với tất cả mọi người.

Một số tổ chức quốc tế để hết tâm trí vào phát triển chuẩn, bao gồm: Tổ chức chuẩn hoá quốc tế ISO (International Organization for Standardization) - có trách nhiệm đối với ngôn ngữ đánh dấu chung chuẩn hoá SGML; IETF (Internet Engineering Task Force) quan tâm đặc biệt đến kiến trúc Internet, tương tác và vận hành Internet. Một trong những chuẩn quan trọng nhất từ viễn cảnh thư viện số là chuẩn tìm kiếm thông tin phân tán Z39.50.

Trong khi các chuẩn tài liệu và thông tin như SGML, HTML, TEI (Text Encoding Initiative), VRML (Virtual Reality Modeling Language) và MARC (Machine-Readable Cataloging) tồn tại, trên thực tế, hầu hết sự trao đổi thông tin điện tử xảy ra qua E-mail, FTP nặc danh, Gopher và các nền tảng trình duyệt Web với TeX, LaTeX, PostScript, PDF, văn bản ASCII và tài liệu định dạng Word và WordPerfect. Hầu hết trong những dạng này không có các cơ chế phân biệt phần đóng góp của nhiều tác giả hoặc nhiều ấn bản, cũng không có khả năng bao hàm các liên kết động tới thông tin khác. Nhiều dạng sử dụng trong thực tế là thương mại, giữ độc quyền và vì thế chúng không có khả năng truy cập phổ quát.

5. Các vấn đề nghiên cứu thư viện số trong tương hỗ ngữ nghĩa

5.1 Tính tương hỗ ngữ nghĩa là thách thức lớn

Cách nhìn chia sẻ là một mạng toàn thể các kho phân tán, trong đó loại đối tượng bất kỳ có thể được tìm qua các tập hợp chỉ số khác nhau. Tương lai gần, các công nghệ phải được phát triển để tìm kiếm trong suốt qua các kho phân tán, điều khiển bất kỳ những biến đổi ở các giao thức và dạng, nghĩa là quan tâm tính tương hỗ cấu trúc. Tương lai xa, các công nghệ phải được phát triển để điều khiển trong suốt những biến đổi ở nội dung và tri thức. Đây là những bước đi theo cách đối sánh khái niệm yêu cầu bởi người sử dụng với đối tượng chỉ số hoá trong kho tài liệu.

Tính tương hỗ ngữ nghĩa sâu sa là khả năng của người sử dụng truy cập nhất quán và rõ ràng tới các lớp đối tượng số và dịch vụ tương tự, phân tán qua các kho không đồng nhất, cùng với tổ chức và dàn xếp bù phần mềm cho những biến đổi theo từng điểm. Để đạt được tính tương hỗ ngữ nghĩa đòi hỏi mô tả bằng tìm kiếm, trao đổi đối tượng và các giao thức tìm kiếm đối tượng. Vấn đề ở đây bao hàm định nghĩa, sử dụng, bắt và tính toán siêu dữ liệu từ các đối tượng, cả văn bản lẫn multimedia, sử dụng mô tả đối tượng tính toán, tổ chức và tích hợp các kho không đồng nhất với ngữ nghĩa khác hẳn nhau, bó chùm và tổ chức phân cấp tự động thông tin, các giải thuật đánh giá tự động, xếp hạng và thẩm định chất lượng, thể loại và các đặc tính thông tin khác.

Định nghĩa và sử dụng siêu dữ liệu, bó chùm và tổ chức phân cấp tự động thông tin là các thành phần chính để xây dựng các hệ thống phân loại tự động đối với thư viện số.

5.2 Nghiên cứu về tính tương hỗ

Các hệ thống phân loại thư viện và các từ điển đồng nghĩa chủ đề riêng biệt như phân loại thư viện quốc hội Mỹ, phân loại Dewey và hệ thống ngôn ngữ y học thống nhất UMLS là những nỗ lực đáng kể của con người để có người quản lý thư viện được huấn luyện giỏi về hệ thống phân loại, gán nhãn tri thức nhất quán. Các hệ thống phân loại thư viện và các từ điển đồng nghĩa thường bắt danh từ/cụm danh từ và chỉ biểu diễn các quan hệ hạn chế. Biểu diễn này thường thô nhưng chính xác.

Các biểu diễn trí tuệ nhân tạo như mạng ngữ nghĩa, hệ chuyên gia và bản thể học phản ảnh cách tiếp cận bắt tri thức khác. Các biểu diễn như thế thường giàu hơn và mịn hơn. Chỉ các nguyên mẫu thực nghiệm trong những lĩnh vực hẹp được tạo ra. Tính hữu ích của chúng trong các ứng dụng thư viện số cỡ lớn vẫn là đáng nghi ngờ.

Cách tiếp cận truyền thống để tạo ra các hệ thống phân loại và nguồn tri thức trong khoa học thư viện và trí tuệ nhân tạo kinh điển thường được xem xét từ trên xuống top-down vì biểu diễn tri thức và dạng được định nghĩa trước bởi các chuyên gia và nhà quản lý thư viện có kinh nghiệm. Quá trình sáng tạo tri thức là có cấu trúc và hoàn toàn xác định. Cách tiếp cận từ dưới lên bottom-up bổ sung để sáng tạo tri thức được đề xuất bởi các nhà nghiên cứu về học máy, phân tích thống kê và mạng nơ-ron.

Dựa vào cơ sở dữ liệu thực, các nhà nghiên cứu phát triển các chương trình phân đoạn và chỉ số hoá tài liệu một cách hệ thống, nhận dạng mẫu trong các cơ sở dữ liệu multimedia khác nhau. Phân tích các cơ sở dữ liệu chứa dữ liệu có cấu trúc và số thường được coi là khai thác dữ liệu/khám phá tri thức. Tạo ra tri thức một cách giải thuật từ các cơ sở dữ liệu multimedia, đặc biệt là văn bản được coi là lõi của quản trị tri thức.

Trong số các kỹ thuật phân tích và chỉ số hoá ngữ nghĩa được coi là có thể mở rộng được, các lớp giải thuật và phương pháp sau đây được khảo sát và thử nghiệm trong thư viện số.

5.2.1 Nhận dạng đối tượng, phân đoạn và chỉ số hoá

Các kỹ thuật quan trọng nhất trong tìm kiếm thông tin bao hàm nhận dạng đặc tính khoá ở đối tượng. Chỉ số hoá tự động và xử lý ngôn ngữ tự nhiên thường được dùng để trích lọc tự động từ khoá/cụm danh từ có nghĩa từ văn bản. Các kỹ thuật chỉ số hoá và phân đoạn dựa vào văn bản, màu sắc và hình dạng thường được dùng để nhận dạng ảnh. Đối với ứng dụng audio và video, nhận dạng tiếng nói và phân đoạn cảnh được dùng để nhận dạng ký hiệu có nghĩa trong luồng audio và video.

Thư viện số phát triển một kỹ thuật phân đoạn danh từ đối với chỉ số hoá tài liệu văn bản. Đối với chỉ số hoá thuật ngữ, chỉ số hoá cụm danh từ để xác định các khái niệm từ một kho tài liệu. Nó bắt đầu với một quá trình mã hoá văn bản để tách biệt ký tự phân cách và các ký hiệu. Nó tuân theo chỉ mục tiếng nói từng phần POST và các luật phân đoạn danh từ ngữ pháp. Đối với thư viện số, kỹ thuật phân đoạn danh từ sản xuất chỉ số chính xác hơn chỉ số hoá từ đảo và trợ giúp tìm kiếm dựa vào nội dung. Bằng cách dùng kỹ thuật xử lý ngôn ngữ tự nhiên mở rộng được, thư viện số có khả năng chỉ số hoá hiệu quả, tự động và chính xác các kho tài liệu của riêng nó.

5.2.2 Phân tích ngữ nghĩa

Một số lớp kỹ thuật được sử dụng đối với phân tích ngữ nghĩa văn bản và đối tượng bao gồm:

- Học máy ký hiệu như không gian ấn bản.
- Thu gộp và phân loại dựa vào đồ thị như thu gộp phân cấp của Ward.
- Phân tích thống kê đa mục tiêu như chỉ số hoá ngữ nghĩa, xác định tỷ xích đa chiều, hồi quy.
- Tính toán dựa vào mạng nơ-ron nhân tạo như mạng lan truyền ngược, ánh xạ tự tổ chức Kohonen và lập trình tiến hoá/lập trình di truyền.

Các kỹ thuật phổ biến này là lựa chọn tốt cho xử lý, phân tích và tóm tắt lượng thông tin multimedia lớn, thay đổi nhanh và khác nhau.

Kỹ thuật không gian khái niệm là một ví dụ về phân tích thống kê, ngữ nghĩa kho tài liệu thư viện số cỡ lớn. Không gian khái niệm được tính toán cho các kho tài liệu có cỡ 100000 trang Web, 1 triệu bản tóm tắt công nghệ và 10 triệu bản tóm tắt y học.

5.2.3 Biểu diễn tri thức

Các kết quả từ quá trình phân tích ngữ nghĩa có thể được trình bày bằng một trong những biểu diễn tri thức sau đây:

- Các hệ thống phân loại
- Các mạng ngữ nghĩa
- Các luật quyết định hoặc logic vị từ.

Nhiều nhà nghiên cứu cố gắng tích hợp các kết quả như thế với các cấu trúc tri thức sáng tạo của con người đang tồn tại như bản thể học, chủ đề và từ điển đồng nghĩa. Sự kích hoạt dàn trải dựa vào các phương pháp suy diễn thường được sử dụng để nghiên cứu kỹ lưỡng các cấu trúc tri thức cỡ lớn khác nhau.

5.2.4 Tương tác người - máy HCI

Một trong những xu hướng chính ở hầu hết ứng dụng thư viện số là đặt trọng tâm vào HCI đồ hoạ thân thiện người dùng. Các trình duyệt dựa vào Web đối với văn bản, ảnh và video làm người sử dụng tăng thêm kỳ vọng về biểu diễn và thao tác thông tin. Những thành tựu ở các ngôn ngữ và nền tảng phát triển như Java, OpenGL, VRML và sự sẵn có của các workstation đồ hoạ cao cấp làm cho trực quan thông tin trở thành một lĩnh vực nghiên cứu nhiều triển vọng.

Những thử nghiệm đầu tiên khẳng định sức quyến rũ đồ hoạ của trực quan 3D, đặc biệt đối với thế hệ Web gần đây. Nói riêng, hầu hết người sử dụng thư viện số có thể biểu lộ các phong cách nhận thức khác nhau và có xu hướng thích 3D hơn. Nghiên cứu HCI nhiều hơn trong ngữ cảnh của thư viện số là cần thiết vì sự phong phú về nội dung và dạng vật tải thư viện số và tính đa dạng trong phong cách và nhu cầu của người sử dụng.

6. Kết luận

Thư viện số là một dạng của công nghệ thông tin, trong đó ảnh hưởng xã hội có tính chất quan trọng như là thành tựu công nghệ.

Thư viện số trở nên quan trọng về mặt quốc gia và quốc tế, một phần là do sự tăng theo hàm mũ của thông tin trên Web.

Công nghệ thư viện số sẽ thống trị Internet của thế kỷ 21. Có một tỷ kho phân tán trên khắp thế giới, trong đó mỗi một cộng đồng nhỏ bảo trì một kho tài liệu tri thức riêng của mình. Chỉ số hoá ngữ nghĩa là có giá trị đối với mỗi một kho, dùng ngữ nghĩa mở rộng nhằm tạo ra trợ giúp tìm kiếm và điều hướng cho hệ thống thuật ngữ chuyên ngành của mỗi một cộng đồng. Sự chuyển qua khái niệm chỉ số hoá ngữ nghĩa tạo khả năng cho thành viên của một cộng đồng dễ dàng tìm kiếm hệ thống thuật ngữ của cộng đồng khác.

❖ Tài liệu tham khảo

[1] A.Barth, M.Breu, A.Endres, A.de Kemp, *Digital Libraries in Computer Science*, Springer, 1998.

[2] B.R.Schatz, Information Retrieval in Digital Libraries, *Science* **275** (1997) 327-334.

[3] I.H.Witten, R.McNab, The New Zealand Digital Library, *The Electronic Library* **15** (1997) 495-504.

[4] Special Issue of JNCA on Digital Libraries, *Journal of Network and Computer Applications*, Vol. **20**, 1997.

[5] K.Fullerton, J.Greenberg, M.McClure, E.Rasmussen, D.Stewart, A Digital Library for Education: the PEN-DOR Project, *The Electronic Library* **17** (1999) 75-82.